UNIVERSIDAD CARLOS III DE MADRID
Escuela Politécnica Superior

Ingeniería de Telecomunicación

Proyecto Fin de Carrera

Análisis y caracterización de técnicas para el Filtrado de Contenidos digitales en el acceso a Internet y las nuevas Tecnologías.

Autor: Mercedes Núñez Mayor
Tutor: Marcelo Bagnulo Braun

Octubre 2015

Ingeniería Telemática

Índice

Índice de ilustraciones

Índice de tablas

Introducción

En los últimos 40 años, la aparición del ordenador ha influenciado en gran medida nuestras vidas. Internet ha significado la gran revolución en la forma de entender tanto las comunicaciones como las relaciones sociales proporcionando un nuevo entorno a un mundo globalizado en constante cambio.

El impacto de Internet así como de las nuevas tecnologías de información en la sociedad actual no siempre es positivo. En este nuevo mundo tecnológico, la información sin control navega a raudales por la red de redes agrupando un gran conjunto de denotaciones y connotaciones, de acuerdo a los grupos de usuarios y a los servicios cambiantes y en continua evolución.

El anonimato que presenta Internet y su mal uso, puede llegar a encubrir muchos contenidos ilícitos e incluso atentar contra los más pequeños. Es en este momento cuando se necesita además de educación, protección. Además la dependencia actual que ciertos dispositivos, como teléfonos móviles, genera en edades muy jóvenes, empieza a preocupar ya que está influenciando al futuro de nuestra sociedad.

En el ámbito laboral, el uso de Internet también es relevante y puede variar desde su aplicación como principal fuente de información a la hora de enfrentarse a nuevos retos o problemas, alcanzando incluso niveles negativos al mediar directamente en la productividad de los trabajadores y por ende de las empresas llegando incluso a afectar la reputación de las mismas.

En este proyecto veremos cómo una solución de filtrado de contenidos puede ayudar a solventar este tipo de situaciones problemáticas derivadas del uso incorrecto de Internet a la par que analizamos y evaluamos desde el punto de vista técnico una solución real.

i. Motivación

En las últimas décadas, la aparición del ordenador seguida de Internet está marcando las formas de comunicación y costumbres de las sociedades en tiempos relativamente cortos.

Bajo un entorno en el que visiblemente Internet se puede situar como una herramienta clave de nuestra sociedad y que continuamente se encuentra en evolución, surge la necesidad de observación y dirección de su uso en los ámbitos que como sociedad debemos proteger y defender para no caer en sociedades descontroladas, demagógicas e individualistas.

Según se ha ido avanzando en el tiempo, los usos de Internet han variado notoriamente. Desde su empleo como buscador de información (sin recaer si los datos son relevantes o no) se puede extender y destacar el empleo de la navegación por la red para comunicarse (local o mundialmente), transferir e intercambiar de archivos, buscar entretenimiento, hacer negocio, teletrabajar…

Este gran abanico de posibilidades en el uso de Internet resulta crítico ya que por desgracia en algunos casos se llega al extremo de generar individuos totalmente ajenos a su realidad y sin contacto social. Suelen ser los más jóvenes, desarrollan trastornos de personalidad con síntomas semejantes a los de la drogadicción al presentar una "adicción informática", como señalan los expertos. Por esto, y por otros temas que actualmente están en boga como la pornografía infantil, muchos padres intentan proteger con las herramientas con las que cuentan a sus hijos.

En el caso de las empresas, las amenazas son muy distintas pero también buscan protección. Para evitar que les afecte negativamente una herramienta fundamental como es Internet en muchos puestos de trabajo, su salvaguardia comienza por establecer una serie de políticas que determinan dónde pueden acceder sus trabajadores y dónde no. Así

persiguen mejorar su productividad e imagen aprovechando al máximo las bondades de Internet.

Teniendo en cuenta que hace años la mayoría de las personas sólo tenían acceso a Internet desde su trabajo o universidad y hoy en día se tiene acceso a la banda ancha incluso desde los teléfonos móviles, el giro ha sido radical. Pero la preocupación surge cuando vemos que hasta los más pequeños son capaces de acceder a cualquier contenido tanto educativo como para adultos o violento desde un inofensivo teléfono o tableta.

Como no es posible regular la información que aparece en Internet, es en este punto donde es necesario dar a conocer algún tipo de solución capaz de ofrecer control sobre el acceso a Internet y cuyo uso sea generalizado cubriendo diferentes entornos y necesidades.

En este proyecto se ha tratado de realizar un análisis de un sistema de filtrado de contenidos debido a las necesidades acaecidas en la sociedad actual. Cuyo uso ofrecería tranquilidad a los padres al proteger la integridad de sus hijos. Y, en el caso de las empresas, eliminaría cualquier tipo de perjuicio ante el uso con fin particular y privado que los empleados pueden originar.

ii. Objetivos

El objetivo principal de este proyecto es evaluar una aplicación de filtrado de contenidos desplegada entre un usuario e Internet con la finalidad de bloquear el acceso a contenidos considerados para adultos (como pornografía, violencia…) mediante una página de bloqueo definida.

Para ello partiremos de las definiciones iniciales del análisis de contenido. Veremos cómo ha ido evolucionando este concepto hasta encajar en la situación actual en la que nos encontramos y al entorno de estudio, Internet.

Las metas que se perseguirán a lo largo de este proyecto serán:

- El estudio de una solución de análisis de contenidos.
- El diseño de un sistema basado en análisis de contenidos y su aplicación en Internet.
- El análisis del comportamiento de una solución de filtrado de contenidos en Internet. Para ello se definirá una plataforma de pruebas, se instalará y configurará la solución y se evaluará su respuesta en términos de rendimiento, latencia...
- Finalizar con un ejemplo de caso de uso real de la solución de filtrado de contenidos.

Queda fuera del ámbito de este proyecto la decisión de los contenidos que deberían ser lícitos o ilícitos. Los casos utilizados en esta memoria serán definidos para ejemplificar en la aplicación de una solución de filtrado en un entorno ficticio.

iii. Resumen

En el primer capítulo se describen los inicios del análisis de contenido y las características bajo las cuales se desarrolló este concepto. También se hace un recorrido sobre el origen de Internet ya que en este proyecto se utilizar este entorno para estudiar una solución de contenidos web.

El segundo capítulo se centra en el ámbito de aplicación del análisis de contenidos y las técnicas sencillas en las que basar el análisis de páginas web. Se finaliza describiendo una arquitectura básica genérica de un sistema de filtrado de contenidos web aplicado a las nuevas tecnologías e Internet.

En el tercer capítulo se describe el entorno de pruebas sobre el que se ha desplegado e instalado un filtro de contenidos básico. Esta plataforma servirá para realizar alguna evaluación básica del sistema.

En el cuarto capítulo, se muestra un caso de uso para una solución de filtrado de contenidos.

En el quinto capítulo, se detalla la planificación del proyecto a partir de la cual se ha realizado el presupuesto general de los costes asociados a él. Incluye un desglose de las tareas realizadas y el tiempo invertido en cada una.

En el sexto capítulo, se presentan las conclusiones finales alcanzadas así como las futuras líneas de trabajo a seguir para completar un solución de filtrado de contenidos digitales en el acceso a Internet y las nuevas tecnologías.

Por último, al final del documento se han incluido una serie de anexos con información añadida como un glosario de acrónimos y términos, la bibliografía y algunos detalles sobre los elementos utilizados durante la realización de las pruebas.

CAPÍTULO 1: Estado del arte

1. Introducción

El mundo de las comunicaciones ha evolucionado de forma muy rápida y en poco tiempo. La tecnología avanza y cada vez está más introducida en nuestra vida diaria con lo que se está fraguando un importante cambio en nuestra forma de relacionarnos y comunicarnos.

En ciertos ámbitos, se hace necesario contar con una defensa que incorpore un sistema de protección tanto para empresas, como para niños o adolescentes.

En este capítulo, veremos la estructura bajo la que se apoyan las herramientas de análisis de contenidos.

2. El origen del análisis de contenidos

Comprender qué es el "análisis de contenido" es el pilar básico sobre el que se desarrolla el fundamento teórico de este proyecto. Es necesario conocerlo para poder determinar tanto la clave de los métodos que se utilizan en el análisis de contenidos web que veremos a lo largo de este proyecto, como el funcionamiento de una solución ideada para analizarlos.

2.1. Concepto

En este proyecto se entiende *análisis de contenido* como la técnica de interpretación de textos que contienen las páginas web que representan un contenido que leído y/o visualizado e interpretado determina algún tipo de conocimiento.

El contenido que se puede visualizar en una página web puede ser textual o visual. La interpretación de un contenido podrá ser objetiva o subjetiva. Sin embargo, la lectura que

se hará de un contenido deberá ser totalmente objetiva y sistemática para poder aplicar una técnica de análisis. Así se deberá entender el análisis de contenido como el resultado de la observación y producción de los datos y la interpretación o análisis de los mismos.

La interpretación del contenido desde el punto de vista de si es adecuada o no queda fuera del propósito de este proyecto.

2.2. Definiciones de análisis de contenido

La interpretación de un contenido viene marcada por el *contexto*. El contexto se verá como un marco que rodea y engloba con información de ayuda al lector para que pueda captar el contenido y el significado de todo lo que se dice en el texto. Con esta explicación, lo que se pretende es hacer ver cómo texto y contexto son dos aspectos fundamentales en el análisis de contenido.

Llegados a este punto, comenzaremos con el acercamiento a las definiciones más características del análisis de contenido y de los distintos elementos que las conforman y el análisis de las mismas que se han ido manejando a lo largo de los años.

En 1952 Berelson defendía que el análisis de contenido era "una técnica de investigación para la descripción objetiva, sistemática y cuantitativa del contenido manifiesto de una comunicación". Para Berelson, el análisis de contenidos debía contar con los siguientes atributos:

- 'objetividad' para que los resultados de los procedimientos generasen siempre el mismo resultado.
- 'sistematización' o pautas ordenadas que incluyesen por completo el contenido observado.
- 'cuantificable' porque se debía poder enumerar la información y medir las características que se querían analizar.

- 'manifiesto' que evitase tener que realizar un análisis profundo que ofreciese resultados no fiables.

En 1969 Hostil y Stone aportarían una definición del análisis de contenido que incluiría varios aspectos muy importantes a la realizada anteriormente por Berelson. Para ellos, el análisis de contenido era una técnica de investigación que creaba inferencias identificando de manera sistemática y objetiva características específicas de un texto. El concepto de inferencia sería tan importante al introducir la idea que los mensajes de los datos podrían ser totalmente diferentes a los mensajes directamente perceptibles.

Pero tan importante es mostrar hechos como interpretarlos, con lo que la cuantificación también sería importante. Y, en muchos casos, en los contenidos escondidos se encuentra el verdadero sentido del texto. Por ello, numerosas discusiones se generarían al respecto durante esta época.

En 1990 Krippendorff definiría "reproductividad" como la convergencia entre los atributos de objetividad y sistematización. Desligándose la idea de que todas las reglas que gobiernan los estudios debían ser explícitas y objetivas para poder aplicarse a cualquier análisis de forma sistemática.

Krippendorff completaría su concepto de análisis de contenido bajo la definición de "técnica de investigación destinada a formular, a partir de ciertos datos, inferencias reproducibles y válidas que puedan aplicarse a su contexto". Así todo análisis de contenidos que se precie, debería generarse en relación con el contexto de los datos sobre el cual se apoya.

La definición que englobaría todas las vistas anteriormente sería la de Laurence Bardin. En el año 1996 definiría el concepto de 'análisis de contenido' como "el conjunto de técnicas de análisis de las comunicaciones tendentes a obtener indicadores (cuantitativos o no) por procedimientos sistemáticos y objetivos de descripción del contenido de los

mensajes permitiendo la inferencia de conocimientos relativos a las condiciones de producción/recepción (contexto social) de los mismos".

De todos las definiciones vistas, la conclusión final de la definición de todo análisis de contenido será la que contenga las técnicas necesarias para explicar y sistematizar el contenido de los datos (textos, sonidos e imágenes) incluyendo la expresión del contenido basado en rasgos cuantificables o no para poder obtener las deducciones lógicas que el autor pretende dar a través del mensaje y el contexto.

2.3. Antecedentes de las técnicas de análisis de contenidos

Para conocer la historia del análisis de contenidos veremos el origen de las primeras propuestas metodológicas y analizaremos la sistematización de sus reglas. Sin olvidar la influencia que la aparición del ordenador al utilizarse como herramienta fundamental del análisis.

2.3.1. Propuestas metodológicas

El entorno que promovió notablemente el desarrollo del análisis de contenidos fue la interpretación de los textos sagrados como los himnos religiosos o pasajes de la Biblia. Pero también, la prensa y los periódicos.

En el siglo XX todos los fenómenos simbólicos se convertían en análisis de contenidos. Se utilizaban cadenas de símbolos, frecuencias de fonemas y palabras de los textos. En la prensa, mediante registros verbales y en la radio, mediante símbolos audibles. Tantos análisis y definiciones cuestionarían la validez de los procedimientos y de los resultados para verificar la fidelidad de los codificadores y medir la productividad del análisis.

Y así se llegaría a la pelea entre el análisis de contenido cuantitativo y cualitativo. Entendiendo el análisis cuantitativo como el método para aportar la información sobre la frecuencia de aparición de ciertas características de contenido. Y el análisis cualitativo

como la ausencia o presencia de una característica de contenido dada, o de un conjunto de características, en un cierto fragmento de mensaje que es tomado en consideración. Se tomaría conciencia de que la función principal del análisis de contenido es la inferencia ya que en síntesis, el análisis de contenidos es un método científico capaz de ofrecer inferencias a partir de datos esencialmente verbales, simbólicos o comunicativos.

2.3.2. Aparición de los programas de ordenador

Con la llegada de las nuevas tecnologías y del ordenador, el campo del análisis de contenido experimentó un notable auge ya que aumentó el interés de la traducción automática, resúmenes y sistemas mecánicos de recuperación.

En 1958 aparecería el primer análisis de contenido por ordenador a partir de rutinas de recuperación de información ejecutadas sobre una serie de leyendas populares. Y, posteriormente, sobre documentos políticos. Pero en estos primeros análisis por ordenador, pesaría mucho la falta de registro de las comunicaciones no verbales y surgiría la necesidad de establecer categorías estandarizadas.

En la década de los 60, aparecería el primer software de análisis de contenidos bajo el nombre de *General Inquirer.* Este programa se basaba en el uso de 'diccionarios' que no eran más que tablas de indexación capaces de marcar y repartir en categorías o subcategorías las unidades del texto y su uso se intensificó durante esta década. A día de hoy, este software se sigue desarrollando en la Universidad de Harvard.

Pero rápidamente aparecerían competidores como el programa *Words System Manual* que ofrecía como ventaja la clasificación de las palabras de acuerdo a las características del texto.

Aunque el boom informático llegaría en los años ochenta con la aparición de los primeros programas específicos basados en el análisis cualitativo como *AQUAD, MAX,*

HIPER RESEARCH, NUDIST, ALTLAS, etc... Los cuales se extendieron de manera muy rápida y aún a día de hoy algunos se continúan desarrollando. Estos programas facilitarían el manejo automático de los datos, el proceso de análisis e interpretación de los mismos y la elaboración de la teoría involucrada en esos datos.

Así poco a poco los nuevos programas de ordenador ofrecían mejoras en los análisis literales con lo que ayudaron a dar un importante empuje a las técnicas de análisis de contenidos.

2.4. Características básicas

El análisis de contenidos requiere la distinción de algunos pasos o elementos a partir del método científico para lograr sus objetivos. A continuación se describen los elementos necesarios previos al proceso de análisis.

2.4.1. Definición del objeto de análisis

Primeramente hay que determinar qué contenido se va a estudiar y por qué es importante. Nos podremos basar tanto en la determinación del problema como en la delimitación del tiempo. Es decir, podremos comenzar seleccionando una dirección, una situación, un hecho, un comportamiento o el espacio, las personas y el contexto donde se decide investigar.

Una vez que se tiene claro el problema a investigar, se deben aclarar los conocimientos previos sobre la materia y encajarla en un marco teórico adecuado. Ya que sin este marco sería muy difícil explicar los resultados que el análisis trata de aclarar.

2.4.2. Definición de la unidad de análisis

Dependiendo del propósito del análisis, se deberá aplicar la técnica adecuada para obtener los datos necesarios en el análisis. Para recabar de manera correcta la información

en forma de datos, se necesita comprender antes qué son los datos teniendo en cuenta que deberán ser representativos de fenómenos reales y presentar durabilidad en el tiempo. Así un dato podría definirse como una unidad de información registrada en un medio duradero, que se distingue de otros datos, puede analizarse mediante técnicas explícitas y es pertinente con respecto a un problema determinado.

Los tipos de unidades de análisis que se pueden distinguir son tres: unidades de muestreo, unidades de registro y unidades de contexto.

Se entiende por *unidades de muestreo* las porciones de observación genéricas a partir de las cuales se obtiene el material que será estudiado para medir la frecuencia de los conceptos definidos. Un ejemplo de unidad de muestreo sería un periódico.

Se entiende por *unidad de registro* la parte de la unidad de muestreo que es posible analizar de forma aislada. En un texto pueden ser palabras, temas (frases, conjunto de palabras), caracteres (personas o personajes), párrafos, conceptos (ideas o conjunto de ideas), símbolos semánticos (metáforas, figuras literarias), etc… Siguiendo con el ejemplo anterior, como patrón de unidad de registro tendríamos la sección de un periódico.

Se entiendo por *unidad de contexto* la porción de la unidad de muestreo que tiene que ser examinada para poder caracterizar una unidad de registro aportando el significado de la misma. A veces la unidad de registro y contexto suelen coincidir pero nunca una unidad de contexto será menos extensa que la unidad de registro en cuanto a porciones de comunicación se refiere. Así nuestro ejemplo sería la portada del periódico.

Es importante decidir cómo se definirá el campo de observación de contenido. Empíricamente se ha comprobado que es mucho más útil realizar un muestreo aleatorio en la toma de unidades de análisis.

No debemos olvidar que al tratarse el análisis de contenido de una técnica cualitativa, el análisis podrá modificarse de acuerdo a la obtención de los datos para mejorar los resultados obtenidos.

2.4.3. Sistema de codificación

Siguiendo con el proceso de análisis, tras el muestreo realizado para obtener sus unidades seleccionando las más pequeñas para el estudio, se deben codificar para poder describirse de forma analizable.

El proceso por el que los datos brutos se transforman sistemáticamente en unidades que permiten una descripción precisa de las características de su contenido se llamará codificación. Y este proceso deberá cumplir los siguientes puntos:

- La presencia o ausencia de los elementos.
- La frecuencia de aparición que aumenta la importancia de una unidad de registro cuanto mayor sea.
- La frecuencia ponderada para medir la aparición de uno o varios elementos y asignarles una mayor o menor importancia sobre los demás. Esta ponderación ha de definirse a priori.
- La intensidad con que se manifiesta un elemento.
- La dirección para establecer un sistema de codificación que refleje el sentido bidireccional del texto.
- El orden de aparición que vendrán establecido según la aparición temporal, importancia o función de las unidades de registro.
- La contingencia entendida como la presencia simultánea en un momento dado de dos o más unidades de registro en diferentes niveles de códigos o de contextos.

2.4.4. Sistema de categorías

La definición de las categorías será un proceso relacionado tanto con el material de análisis como con la formación y marco teórico de la investigación.

Al realizar una clasificación de elementos en categorías, se está imponiendo la búsqueda de lo que cada uno de ellos tiene en común con los otros. Este agrupamiento determina la parte que comparten pero también es posible que diferentes criterios influyan en otros aspectos por analogía modificando quizás considerablemente la distribución anterior.

En el proceso de categorización podremos diferenciar dos etapas. En la primera se buscará aislar o inventariar los elementos y en la segunda distribuirlos o clasificarlos de acuerdo a lo que se pretenda hallar. Este proceso, como era de esperar, deberá seguir una serie de reglas para poder realizarse de forma correcta.

Por ejemplo, a la hora de realizar la categorización se podrían tener en cuenta algún punto tal como:

- Cada serie de categorías ha de construirse de acuerdo con un criterio único.
- Cada serie de categorías ha de ser exhaustiva definiendo todos los contenidos.
- Las categorías tienen que ser significativas.
- Las categorías tienen que ser claras evitando cualquier tipo de ambigüedad.
- Deben de ser replicables para asegurar que siempre se pueda realizar la distribución de elemento según el plan de categorización inicialmente propuesto.

2.4.5. Inferencia

El término inferir significa explicar o deducir lo que hay en un texto. En el análisis de contenidos se buscan conclusiones o extraer inferencias/explicaciones embebidas en el texto de forma explícita o implícita.

De acuerdo a Krippendorff, las inferencias que se pueden extraer de un texto comunicativo son numerosas y entre ellas se pueden encontrar:

- *Sistemas.* Se pueden inferir distintos sistemas extrayendo conocimientos sobre sus componentes, sobre las relaciones internas y sobre las transformaciones.
- *Estándares:* Se puede evaluar la calidad, nivel, neutralidad y objetividad de un escrito. Así como inferir su calidad o defectos, proximidad o lejanía respecto a un criterio determinado, comprobando si se alcanza o no dicho criterio.
- *Índices:* La determinación de estándares pueden ir acompañada de la búsqueda de indicadores o huellas para medir los contextos tales como el tema de un escrito.
- *Comunicaciones:* Las marcas de los escritos como citas, supuestos o alusiones… pueden proporcionar información semejante a un intercambio de opinión directo del texto.
- *Procesos institucionales*: A través de los que se puede inferir cierta información como una aclaración de una postura, por ejemplo, ideológica.

Pero el número de inferencias posibles es tan grande como las proporcionadas por otros análisis como el de una encuesta estadística.

2.5. Fiabilidad

Un dato es fiable cuando permanece constante en todas las variaciones del proceso analítico.

La importancia de la fiabilidad proviene de la seguridad de mostrar que los datos han sido obtenidos con independencia del suceso, instrumento o persona que los ha tomado y estudiado.

Volviendo a Krippendorf, se exige que como mínimo dos codificadores describan de forma independiente un conjunto extenso de unidades de registro en los términos de un

lenguaje común como, por ejemplo, un esquema de clasificación de códigos y categorías.

Así, la fiabilidad se expresaría en función del grado de convergencia sobre la asignación de cada una de las unidades a las diversas categorías. Si todos coinciden, se garantiza la fiabilidad pero si el nivel de discrepancias es elevado, se podría considerar que el grado de fiabilidad fuese aproximadamente nulo.

2.6. Tipos de modelos

Acabamos de ver cómo el concepto de análisis de contenido ha ido tomando forma.

Y sería tras la Segunda Guerra Mundial, éste resurgiría con gran fuerza. A partir de este momento aumentaría considerablemente su frecuencia de utilización, sobre todo, con la aparición del tratamiento informático de los datos.

Berelson lo reduciría al contenido "manifiesto" y del análisis "cuantitativo" de los textos, definiéndolo como "una técnica de investigación para la descripción objetiva, sistemática, cuantitativa del análisis del contenido manifiesto de una comunicación".

Krippendorf, por su parte, abogaría por identificarlo como un sistema de "extraer inferencias", y plantearía que "es un procedimiento para extraer inferencias respecto a los emisores y los receptores de la evidencia en los mensajes que se intercambian entre sí".

Bardin, sin embargo, insistiría fuertemente en que no sólo se trata de un instrumento, sino de un conjunto de instrumentos metodológicos diversificados y cada vez más elaborados que se aplican a textos (discursos, frases, documentos escritos, entrevistas, textos literarios, temas de publicidad, diarios, preguntas abiertas en una encuesta, relatos de vida, etc). Y que todo análisis de contenido que se precie debería estar compuesto tanto por el "contenido" del mensaje, como por el "continente" comprendiendo así las comunicaciones más allá de sus significaciones primeras.

Así aunque existen diversas formas de realizar un análisis de contenidos, básicamente se utilizarán los tres modelos que siguientes:

- *Análisis de contenido por temas*

Este análisis sólo considera la presencia de términos o conceptos, con independencia de las relaciones surgidas entre ellos.

En la actualidad ciertos programas de análisis de contenido, como *Anatex*, desarrollado por Raymond Colle (1988), permiten la creación de palabras registradas junto con su contexto (*KWIC:Key-word in context*). Con este sistema se seleccionan determinadas palabras y se obtiene para cada una la trascripción de la oración o parte del texto en la cual aparece, pudiéndose discriminar y reagrupar los significados. Esta técnica tiene la ventaja de que se pueden buscar cadenas de caracteres de diferente amplitud desde raíces (comunes a varias palabras) hasta palabras compuestas o frases enteras.

- *Análisis de contenido semántico*

Este tipo de análisis pretende ante todo estudiar las relaciones entre temas tratados en un texto. Para ello se han de definir los patrones de relaciones que se tomarán en cuenta.

- *Análisis de contenido de redes*

Por último, este tipo de análisis se centra en la ubicación relativa de ciertos componentes al entender que la ocurrencia de los elementos léxicos conlleva una historia textual.

3. El origen de Internet y de la navegación

En este apartado veremos los inicios de Internet y cómo ha evolucionado según las inquietudes de los internautas.

3.1. Comienzos de Internet

El envío de correos electrónicos, relaciones sociales, compras e incluso las operaciones bancarias son tareas que se realizan a día de hoy de forma virtual. Pero muchas más cosas que antes nos parecían un sueño se pueden realizar con la ayuda de Internet.

Para entender cómo se ha llegado hasta este punto, hay que retroceder a la mitad del pasado siglo XX. En esta época, los ordenadores sólo podían realizar una única tarea que se conocía como *batch processing* o procesamiento por lotes, que además era bastante ineficiente. Físicamente no eran cómo nos los imaginamos ahora sino que además crecían en tamaño. Había incluso que alojarlos en lugares específicos con estrictos controles de temperatura donde los desarrolladores debían desplazarse para poder trabajar.

En este entorno, los americanos bajo su ideal de ser los pioneros y líderes en tecnología motivarían una serie de iniciativas desarrollando técnicas de acceso remoto. Tras esta iniciativa, nacería el concepto de compartición de recursos en tiempo real o *time sharing.* Muy interesante ya que era una época en la que los conocimientos los tenían las personas por lo que eran las únicas que podían transferirlos.

Existen muchas creencias sobre el origen de Internet, pero la idea más extendida es que nació hace más de 50 años como un proyecto de investigación en redes de conmutación de paquetes dentro de un ámbito militar americano. Debido a las guerras que se llevaban a cabo por aquellos años, se temía que las comunicaciones por radio no se sustentaran ante un posible ataque al poder verse afectadas de forma muy significativa. Así el Departamento de Defensa Americano a través de su agencia de proyectos de investigación avanzados, conocida como ARPA, instó a que universidades y empresas privadas se unieran a dicho proyecto con la finalidad de mejorar la iniciativa inicial basada en la red telefónica conmutada o RTC. Se encontraban ante la significativa vulnerabilidad de las comunicaciones únicas y limitadas entre nodos centrales, la caída de uno de ellos.

Por el año 1969 surgiría el proyecto denominado ARPANet con el fin de lograr una comunicación garantizada cuya iniciativa buscaba un sistema de comunicación basado en los puntos siguientes:

- *Fiabilidad***,** un sistema que siempre funcionase independiente del estado de la red a través de la cual se estableciera la comunicación.
- *Sencillez en la distribución de la información,* gracias a la nueva cabecera que se incluía en todo paquete que viajase por la red con los datos necesarios para su entrega con total independencia del camino seguido.
- *Mejoras en la transmisión,* ya que al tener un origen militar, otra preocupación que también estaba presente era asegurar la confidencialidad de los datos enviados incorporando técnicas de encriptado.

Para ello, ARPANet utilizaría nodos IMP (*Interface Message Processor*) para controlar las actividades de la red proporcionando el acceso a los ordenadores centrales de la misma. El acceso a los nodos estaría a su vea controlado por otros que se interconectarían mediante el protocolo de red NCP (*Network Control Protocol*) creando así una subred entre ordenadores. Protocolo que posteriormente sería desbancado por el TCP (*Transfer Control Protocol*) mucho más eficiente y que además incluía la funcionalidad de la verificación de transferencia de archivos.

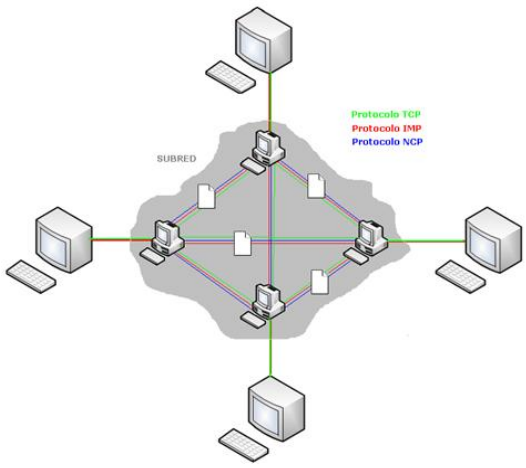


Ilustración 1. Inicio de los protocolos de comunicación

Ideas similares se llevaron a cabo en otros países bajo proyectos científicos como Cyclades en Francia o comerciales como NPL en Inglaterra las cuales impulsaron ciertas características de Internet que actualmente concebimos como fundamentales. Éstas serían:

- El sistema inglés NPL contribuiría con el concepto de conmutación de paquetes al implementar la fragmentación de la información en unidades de tamaño determinado y más pequeño, llamados paquetes. Cada paquete era encaminado hacia su destino final gracias a la información que almacenaba en su cabecera como el destino, origen… sin especificar el camino a seguir por cada paquete ya que este vendría definido de acuerdo a los nodos intermedios de paso. Así, y descentralizando las redes, en el caso de que uno o varios nodos estuviesen caídos, el resto de elementos de red serían capaces de encaminar automáticamente los paquetes de tal forma que no se perdiera en ningún momento la comunicación.
- Por su parte, la red Cyclades sería la responsable de la interconexión entre las diferentes redes existentes en el mundo. Pero aquí surgirían los primeros problemas de compatibilidad debido a la discrepancia en los protocolos de comunicación utilizados en cada una de las redes de área local que posteriormente solventaría la ISO al estandarizar el protocolo en uso TCP con el actual protocolo TCP/IP. Así se garantizaría la comunicación entre todos los ordenadores de la red proporcionando un sistema independiente de intercambio de datos entre ordenadores y redes locales de distinto origen conservando las ventajas relativas a la técnica de conmutación de paquetes.

Al estandarizar las comunicaciones, estaba naciendo la 'red de redes'. Así todas las redes se podrían unir mundialmente al desarrollarse una asociación de miles de redes conectadas entre sí. Se comenzaba a vislumbrar la posibilidad de que un ordenador de una red podía intercambiar información con otro situado en una red remota. Este hecho se

vería fuertemente impulsado por la inclusión de numerosas universidades, centros de investigación, empresas privadas, organismo públicos y demás asociaciones de todo el mundo.

Pero el boom de Internet podría haberse debido a la notable facilidad de uso de los servicios ofrecidos tales como el acceso remoto, el correo electrónico o la aparición de la *World Wide Web (www)* como servicio de consulta de documentos hipertextuales. La definición de una herramienta para controlar el uso de la navegación centrará el marco de este proyecto.

3.2. Evolución del uso de Internet

Inicialmente Internet tenía como claro objetivo el acceso a ordenadores remotos, pero fueron apareciendo nuevas funcionalidades entre las que destacaba por encima de todas la navegación en busca de información.

Actualmente, hoy es más probable perderse en la red, debido al gran abanico de posibilidades que brinda. Además el progresivo avance del uso de Internet ha desbancado al resto de medios convencionales de comunicación como libros, enciclopedias, televisiones, radios, periódicos…e incluso ha llegado a ser la herramienta básica de muchos puestos de trabajo como ya se comentó anteriormente.

A la vista está que Internet no sólo ha evolucionado con respecto a las comunicaciones sino también con respecto a los servicios que ofrece. Si bien al principio sólo permitía ejecutar programas de forma remota, en 1972 se introdujo el *correo electrónico***.** Este nuevo sistema, se convirtió en la actividad de la red que más tráfico generó de forma totalmente sorprendente. Se instauró como sistema básico de comunicación de Internet al hacer posible que cualquier persona desde el lugar en el que estuviera se pudiera comunicar de una manera rápida y barata con cualquier otra persona en cualquier parte del mundo simplemente con que ambas tuvieran una cuenta de correo. Este servicio ha

evolucionado de tal forma que ha llegado a fusionarse con otros servicios de Internet como la mensajería y las redes sociales (ya que muchos portales de Internet disponen de una *página web o webmail* a través de la cual pueden recibir y enviar mensajes) que se siguen utilizando en gran medida en la actualidad.

Otra aplicación que también nació en los comienzos de Internet fueron los extendidos chats más conocidos por entonces como *IRC o Internet Relay Chat*. Estas aplicaciones permitían la comunicación simultánea y en tiempo real con personas que se conectaban al mismo tiempo y desde cualquier parte del mundo. La evolución de estos programas ha sido tan sobresaliente que actualmente permite utilizar la voz, videoconferencia e incluso ha derivado en aplicaciones laborales utilizadas para presentaciones remotas que incluyen audio y video del PC del presentador.

Ante el auge de los chats, muchos proveedores de correo electrónico los incluyeron en sus páginas web o desarrollaron aplicaciones específicas con la idea de que todas las personas que tuvieran contratadas con ellos una cuenta de correo contaran con la funcionalidad de poder chatear o enviarse ficheros en tiempo real.

Pero el servicio más demandado data del año 1989 y es a partir del cual han surgido otros muchos, es el *Word Wide Web* o páginas de la *www.* Podríamos decir que a través de ellas, Internet cuenta con la mayor base de datos del mundo en un soporte innovador ya que estos millones de páginas están llenos de contenidos de todo tipo, repartidos por los miles de servidores que hoy en día se encuentran desplegados por todo el mundo. Estas páginas permiten el acceso a multitud de información multimedia a través de un sistema interactivo de enlaces o links entre diferentes páginas web. De esta manera, los usuarios pueden consultar información, periódicos, buscar trabajo o casa, acceder a las cuentas bancarias, jugar en red… y la red ofrece soluciones a todos según las necesidades diarias.

Cada página web tiene una dirección *URL o Uniform Resource Locator* que la identifica.

El lenguaje utilizado para la composición y edición de estas páginas es el *HTML* o *HyperText Markup Language* de tal forma que cualquier persona podría tener su página web siempre que previamente haya reservado algún espacio web en uno de los servidores que hay por todo el mundo. Un ejemplo sería, *http://www.uc3m.es*

El mundo de las páginas web abrió un gran abanico de posibilidades a nuevos servicios y aplicaciones. Así dentro de este ámbito, aparecieron los buscadores y las redes sociales.

Los buscadores serían los sitios más frecuentes y utilizados en Internet donde obtener información de miles de millones de sitios gracias a sus bases de datos. Referente a las redes sociales, experimentaron un gran auge en los últimos años, sobre todo, con la introducción de los más pequeños de la casa y jóvenes. Su utilización saltaría del ámbito de las relaciones personales a las profesionales. Pero estas redes son el ejemplo más claro de que Internet también puede ser utilizado de forma negativa ya que muchos pederastas se aprovechan engañando a los jóvenes para conseguir sus fines.

Pero otros servicios que también abundan y que surgieron con el nacimiento de Internet son los *grupos de discusión o news* y la *transferencia de ficheros.* En el primero, la gente se subscribía a una lista de distribución de acuerdo al tema específico que les interesaba para recibir información actual relacionada con él. Por otro lado gracias a la *transferencia de archivos,* los usuarios mediante algún programa o protocolo podrían conectarse a un ordenador remoto y descargarse ficheros como, por ejemplo, de música o video.

Las cada vez mayores posibilidades y nuevos servicios que se ofrecen a través del ordenador o móvil están empujando su utilización, sobre todo, entre un segmento de audiencia joven donde las relaciones personales, la descarga de música y los juegos son piezas fundamentales. Pero está claro que la incorporación de tantos individuos a la red implica una mayor cantidad de relaciones virtuales entre personas, soluciones a problemas… Por lo que Internet, podría definirse como una revolución.

3.3. Evolución de usuarios de Internet

Tras la aparición del *WWW*, el incremento en el número de usuarios que utilizan Internet ha sido destacable. Sin embargo, el crecimiento ni ha sido homogéneo ni lineal y el objetivo de su uso también ha cambiado a lo largo de los años.

Partiendo del año 1998, desde las pioneras tierras norteamericanas, comenzaría la extensión de Internet por el resto de países. Así en países como Suecia, Australia, Nueva Zelanda e Islandia se vería cómo comenzaba tímidamente esta nueva andadura tecnológica. Sin embargo, y contrariamente, en los países del continente africano apenas se registrarían usuarios.

Durante el siguiente año, el uso de Internet se extendería rápidamente y con fuerza desde Norte América a Europa. Asia, Japón y Australia alcanzarían el nivel de penetración que presentaba países precursores como Nueva Zelanda. En Oriente Medio, los Emiratos Árabes comenzarían su andadura.

Y sería en este año 1999, cuando se establecieron los dominios con más servidores localizados a los siguientes TLDs: com, net, edu, jp, uk, mil, us, de, ca y au.

Sin embargo, no será hasta el año 2000 cuando Internet se establezca fuertemente en Europa mientras que por otros lugares como en Malasia, Corea del Sur y Chile estarían surgiendo los denominados *hotspots* (o puntos de acceso a Internet mediante redes inalámbricas).

En cuanto al continente africano, los cambios no serían destacables con un uso bastante limitado.

En el año 2001, España comenzaría a ponerse en línea con el resto de países europeos. En el continente africano, Sudáfrica alcanzaría a otros países americanos como Argentina y México.

En el año 2002, Internet comenzaría su entrada más fuerte en Europa por los países del Norte. Otros como Brasil y Turquía, comenzarían sus andaduras conjuntamente con Venezuela y Costa Rica. Mientras que en África, la República Democrática del Congo registraría unos 50.000 usuarios de Internet pero nada comparado como el uso de Internet de otros continentes.

Durante el año 2003, China crecería desde los 20 millones de usuarios de Internet hasta alcanzar la cifra de 79 millones, aproximadamente un 6% de su población total.

Países como Rusia e Irán continuarían creciendo a la par, mientras que en el continente africano, Zimbabue se pondría al mismo nivel que su vecina Sudáfrica.

Pero será el año 2004 el que se recuerde como aquél en el que todo comenzó y en el que Internet experimentaría su gran auge. En ello influiría la aparición de los bloggers, el lanzamiento de Gmail con espacio de almacenamiento gratis por parte del gigante Google que también compraría otras empresas de servicios como Picassa. Por otro lado, Microsoft lanzaría su servicio de noticias sociales llamado Newsboot… Y sería el año del tímido comienzo de la mayor red social, Facebook.

En Europa se registraría que casi la mitad de usuarios residenciales y que 9 de cada 10 usuarios de empresa usaban Internet. En Rusia el nivel de penetración alcanzado sería del 50% de usuarios, duplicándolos Irán y triplicándolos Marruecos.

Y durante el año 2005, continuaría el fantástico crecimiento de Internet. Sería la era del podcast gracias a Apple. Y Facebook se abriría a los colegios y universidades...

Todo ello impulsaría el crecimiento en el número de usuarios incluso en países en los que antes no destacaba. Egipto y Túnez se equipararían a Marruecos. Y en Asia, el mayor nivel de penetración se registraría en la India.

En el año 2006, el titán de Google combinaría todos sus servicios en una plataforma

universal ofreciendo una cantidad ilimitada de espacio de almacenamiento y ancho de banda para almacenar y compartir todo tipo de medios. Pero Facebook, y ante su flamante éxito, se abriría mundialmente comenzando su estrellato.

Argelia, Sudán y Kenia alcanzarían los niveles de Sudáfrica. Países como Mongolia superarían el nivel de penetración de China y Brasil superaría a Reino Unido en el número de usuarios.

El crecimiento de Internet se asentaría en el este de Europa durante el año 2007 a excepción de países como Albania y Ucrania. Por otro lado, en América del Sur iría creciendo paulatinamente como en el caso de Bolivia.

Tras el auge de Internet, muchos periódicos dejarían de editarse en papel durante el año 2008, para salir en versión digital lo que demostraría que hasta este año no habrían aceptado lo que Internet había influido sobre ellos.

A la vista de la descripción anterior, tanto África como el sur de Asia se encontrarían bastante retrasados en cuanto a los avances de Internet. Por ejemplo, en la República Democrática del Congo tan sólo el 0.45% del total de su población sería usuarios de Internet.

En el año 2009 el nivel de penetración se encontraría bastante consolidado. Así las regiones con mayor número de internautas serían Norte América seguida de Australia, Europa y el Sur de América por este orden.

El Medio Oriente, el sur de Asia y África se encontrarían en niveles inferiores siguiendo la línea de años anteriores.

Pero desde el 2010 hasta la actualidad, Internet se ha afianzado mucho en nuestras vidas. Para la vida social recurrimos a Facebook y a los blogs, para la vida laboral a Linkedlin, para el entretenimiento a Youtube, para las compras a eBay, el correo

electrónico se utiliza como sustituto a la hora de escribir una carta o enviar un fax…. Y es que a lo largo de todos estos años los ordenadores pasaron de ser una herramienta para crear contenidos a una herramienta para comunicarse y/o consumir contenidos.

La revolución que estaba causando Internet conseguiría que todas las personas de un planeta estuviesen conectadas entre sí.

De acuerdo a los datos proporcionados por la *ITU*, el organismo especializado en regular las telecomunicaciones a nivel internacional, se ha generado la siguiente gráfica que muestra cómo ha evolucionado el número estimado de usuarios de Internet en los últimos 15 años:

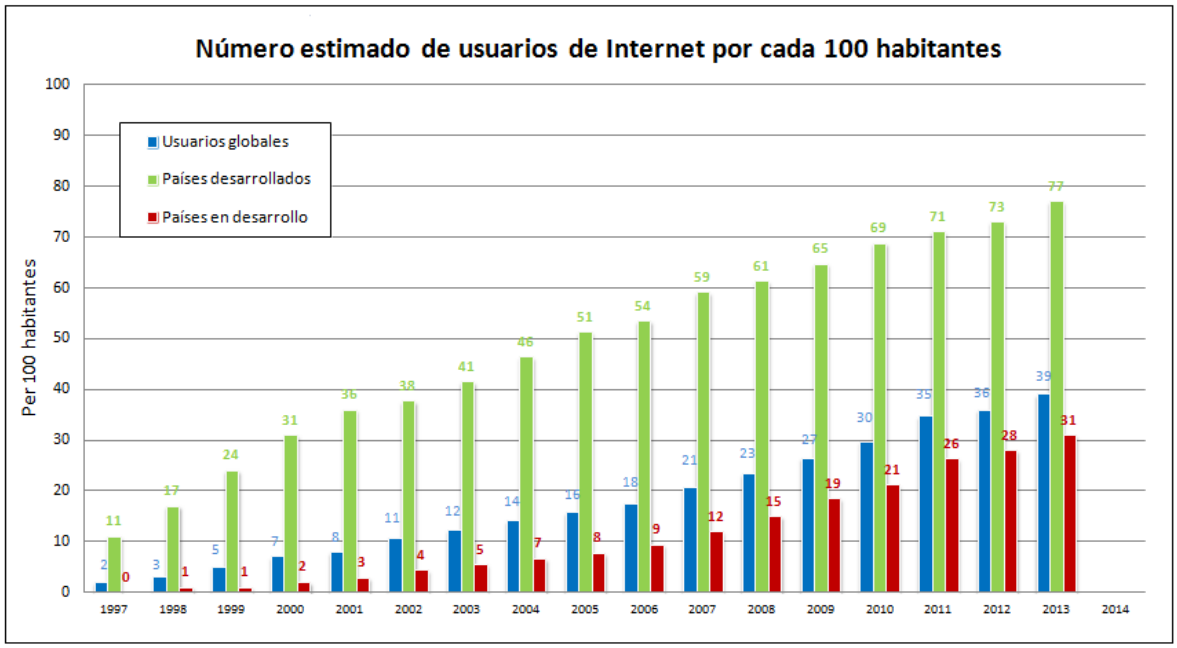


Ilustración 2. Estimación de usuarios de Internet por cada 100 habitantes*

*Datos numéricos obtenidos de estadísticas de la *ITU* (http://www.itu.int/ict/statistics , http://www.itu.int/ITU-D/ict/definitions/regions/index.html)

A partir de datos estadísticos de los usuarios globales de Internet ofrecidos por la ITU para el intervalo de años comprendido entre 1996-2013, podemos aplicar una interpolación polinómica y generar el siguiente gráfico:

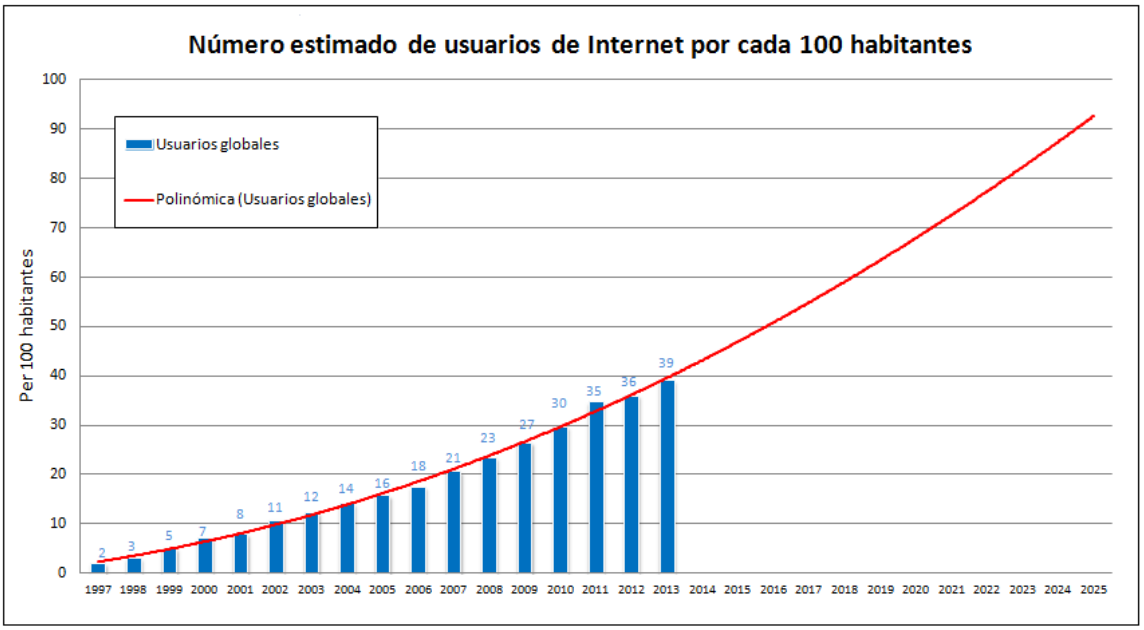


Ilustración 3. Aproximación polinómica de los usuarios globales de Internet

De este gráfico podríamos deducir que en poco más de 10 años la tendencia será que la gran mayoría del mundo hará uso de Internet.

Pero durante esto últimos años, el lugar de acceso a Internet también ha ido variando. Al comienzo, el primer sitio más utilizado para acceder era el lugar de trabajo pero ahora cada vez más internautas se conectan a Internet desde casa y no mayoritariamente desde un ordenador. Las tabletas y móviles han ido ganando mucho terreno.

También el perfil del usuario que se conectaba a Internet ha cambiado con el tiempo. Inicialmente eran exclusivamente adultos pero ahora hasta un niño, que apenas sabe leer o escribir, es capaz de utilizar una tableta para ver sus dibujos favoritos.

CAPÍTULO 2: Diseño de un sistema de análisis de contenidos de Internet

Partiendo de las diversas teorías que surgieron a lo largo del siglo XX a partir de las cuales comenzó la teoría de análisis, hay que tener en cuenta que las nuevas tecnologías de la información y la comunicación han creado nuevos ambientes de aprendizaje que están en constante transformación y que ajustarán las técnicas de análisis de contenido originales.

Este proyecto está orientado en soluciones que protejan a usuarios finales de Internet de la delincuencia en la red, a los menores de la pornografía y el acoso escolar en la red y a las empresas de los actos inadecuados de sus trabajadores en el uso diario de Internet. Así se diseñará un sistema bajo la idea que se pueda construir un entorno seguro de navegación reduciendo las numerosas amenazas existentes en la red.

Durante el resto del proyecto, cuando hablemos de análisis de contenido o análisis a secas, nos centraremos en el marco de navegación de Internet y de las redes de comunicación. Así de acuerdo a las características vistas en el capítulo anterior, la unidad de análisis vendrá definida por:

- Unidad de muestreo: página web
- Unidad de registro: texto de la página web
- Unidad de contexto: palabras (inicialmente)

Tomaremos un modelo basado en un análisis de contenido temático, por lo que será necesario recurrir a categorías de contenido.

En este capítulo veremos las bases en las que se fundamentaría un análisis de contenidos web partiendo de las diferentes técnicas que se pueden aplicar y su uso. Se concluirá con el diseño sencillo de un sistema de análisis de filtrado de contenidos web.

1. Introducción a las técnicas de clasificación

Hemos visto como el análisis de contenidos fue surgiendo tímidamente durante el s.XX con las primeras escuelas de periodismo. Y que no sería hasta mediados de siglo cuando Berelson, Bardin y compañía le darían el impulso necesario que hasta el día de hoy sigue presente en sectores como el de las ciencias humanas.

Vimos que son múltiples las clasificaciones que se hacen cuando se trata el tema del análisis de contenido. En este proyecto, el análisis de contenidos no es un método de obtención de información sino de tratamiento de información y su finalidad será la obtención de datos: objetivos, susceptibles de medición, significativos y generalizables.

Este análisis de contenido cualitativo permitirá determinar la temática de un recurso alojado en la web.

1.1. Agrupaciones de páginas

La forma más sencilla de aplicar el filtrado de contenidos se basa en utilizar categorías temáticas. Para ello habría que definir de forma unívoca cada una de las categorías dentro de las cuales podrá categorizarse una página web por lo que será muy importante que la definición de las mismas se realice de forma sencilla y bien estructurada.

También habrá de determinar, el número de categorías en base al ámbito en el que se busca aplicar el análisis de contenido. Para este proyecto, un diseño de categorías de contenido podría ser:

- **Actividades criminales:** sitios web con contenidos que informan sobre actividades criminales o ilegales. Dentro de este ámbito se incluyen detalles sobre cómo cometer asesinatos, suicidios, sabotajes, preparación de bombas, rotura de cerraduras u otras técnicas para realizar robos.

- **Armas:** sitios web con contenidos que informan sobre la venta de armas de fuego y armas blancas tanto de uso militar, como deportivo o de caza. También se incluyen sitios relacionados con artículos para artes marciales y defensa personal como puños americanos.

- **Compras on-line:** sitios web a través de los cuales pueden realizar la compra venta de productos y servicios.

- **Drogas:** sitios web con contenidos que fomentan el consumo de drogas o facilitan cómo obtenerlas. También estarían incluidos dentro de esta categoría las páginas web que venden medicamentos sin receta médica.

- **Hacking:** sitios web con contenidos explícitos sobre cómo acceder ilícitamente a redes de ordenadores, dispositivos ajenos o la manipulación de los mismos.

- **Juegos:** sitios web en los que se puede jugar en línea o desde los cuales pueden descargarse videojuegos.

- **Niños:** sitios web con contenidos diseñados específicamente para niños con edades comprendidas entre los 3 y los 16 años.

- **Mensajería instantánea:** sitios web que informan o permiten descargarse programas para comunicarse en tipo real con usuarios que previamente han sido invitados.

- **Pornografía**: sitios web con contenidos pornográficos, obscenos o eróticos.

- **Radio y TV por Internet:** sitios web que retransmiten programas de radio o televisión en tiempo real.

- **Redes sociales**: sitios web dedicados a las comunidades en línea donde los usuarios comparten información entre sí. Estos sitios pueden tener propósitos profesionales o de ocio sin incluir los sitios dedicados a contactos.

- **Software ilegal**: sitios web que contienen información o permiten descargar programas tales como música o películas distribuidas de forma ilegal. También se incluyen las páginas que alberguen licencias ilegales de programas software**.**

- **Spyware:** sitios web que contienen software cuyo fin es recoger información de un dispositivo para enviarla a un tercero a través de la red sin autorización del propietario del dispositivo.

- **Violencia:** sitios web con contenidos violentos o que incitan a la misma.

Estos ejemplos de categorías han sido tomadas de soluciones de filtrado que actualmente se encuentran en el mercado como McAffe, Allot, BlueCoat o Palo Alto Networks.

Al utilizar una agrupación de páginas web mediante categorías de contenido, implica categorizar no sólo las miles de millones de páginas web ya existentes sino a estar continuamente categorizando las nuevas páginas web que todos los días nacen. La distribución de los elementos, páginas web, debería realizarse de forma manual para disminuir el número de falsos positivos.

1.2. Análisis del contenido de un sitio web

Pero como es imposible tener categorizadas todas las páginas web existentes y categorizar las páginas nuevas según van naciendo, es necesario implementar otra serie de técnicas para que en el caso de encontrarnos con una página web que no esté contenida en una categoría se pueda formalizar el proceso de clasificación.

Lo ideal sería encontrar un algoritmo para la clasificación automática de los contenidos de las páginas web. En primera aproximación podríamos pensar en realizar un análisis lexicográfico de la página web para posteriormente complementarlo con uno sintáctico. Pero para poder establecer una inferencia de clasificación a partir del análisis lexicográfico, se necesita decidir cuál será la unidad de análisis que se va a utilizar y que deberá estar previamente categorizada.

Teniendo en cuenta que el análisis de contenido se va a realizar para peticiones de contenidos de páginas web, es importante que este análisis sea rápido para no penalizar la experiencia de usuario. En vez de buscar palabras completas será más eficiente, desde un punto de vista computacional, buscar patrones o raíces de palabras. Luego las unidades de análisis ya están caracterizadas, serán patrones de palabras y sólo habrá que decidir en base a las coincidencias encontradas el contenido de la página web. Para poder hacer esto, habrá que definir también algún mecanismo que determine cuánto de probable es que dicha unidad de análisis pertenezca a una o a varias categorías de contenido. Apareciendo así el concepto de diccionario para agrupar las unidades de análisis por categoría y el de peso para determinar el grado de pertenencia de la unidad de análisis a una categoría de contenido.

Lo más importante es definir bien los diccionarios ya que son la herramienta fundamental para la detección de textos en base a categorías. Para generarlos se pueden utilizar clasificadores bayesianos y/o redes neuronales ya que permiten procesar y analizar la información para encontrar patrones repetitivos a la vez que dan la posibilidad de realizar una ponderación de la posibilidad de ocurrencia de los mismos. Mediante estas técnicas se puede modelar la estructura de las categorías y las relaciones con las unidades de análisis. Al aplicar este tipo de algoritmos para la predicción es necesario tener en cuenta que también influirá notoriamente en el resultado la calidad de los datos de entrenamiento.

En nuestro caso, se podrán utilizar como conjuntos de entrenamiento agrupaciones de URLs cuya naturaleza es conocida, es decir, se conoce a priori si pertenecen o no a la categoría bajo estudio. De esta forma, a la salida del entrenamiento se obtendrían unas reglas basadas en los conjuntos de entrenamiento que posteriormente permitirían clasificar los contenidos. Estas reglas serán el diccionario con las unidades de análisis y sus pesos.

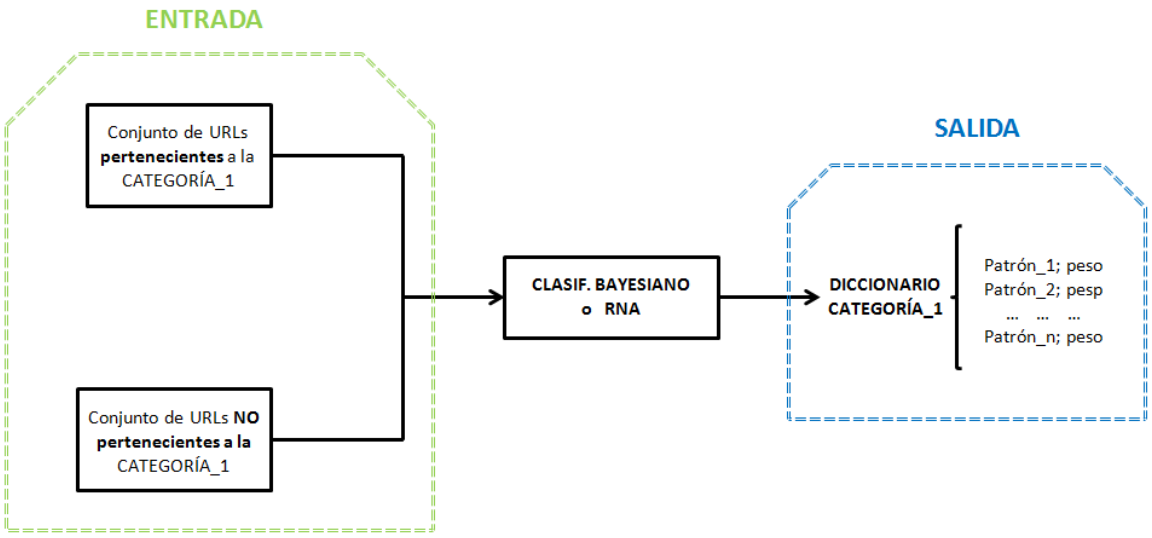


Ilustración 4. Entrenamiento

Una vez que ya está conformado el diccionario, al realizar un análisis de contenido hay que contabilizar los pesos registrados por cada una de las unidades de análisis incluidas en un diccionario y comparar si el peso total supera un valor umbral definido. En el caso que el peso total observado sea mayor que el umbral, la detección sería positiva y el contenido analizado pertenecería a esa categoría. Y en el caso que el peso total observado fuese menor que el umbral, la detección sería negativa y el contenido analizado no pertenecería a una categoría.

2. Aplicación de las técnicas de clasificación

Las técnicas vistas en el capítulo anterior se pueden utilizar para definir las diferentes formas de analizar un contenido. A continuación veremos algunas de sus aplicaciones y usos posibles en sistemas de filtrado de contenidos.

2.1. Categorías de contenidos

Agrupar los sitios web en categorías sería la forma más eficiente de implementar soluciones de filtrado de contenidos.

El diagrama de análisis de contenidos basado en categorías sería el siguiente:

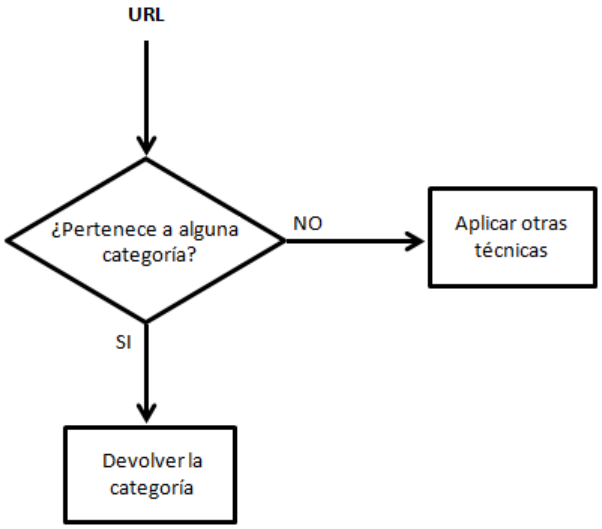


Ilustración 5. Diagrama del análisis de categorías

2.2. Sistemas de clasificación normalizados

La preocupación por el acceso a la información que los menores podían hacer a través de Internet ha preocupado a diferentes organismos mundiales. Esto originó que se crearan diversas iniciativas que describiesen el contenido de un sitio sin necesidad de tener que acceder a él.

Pero estos sistemas de clasificación no pueden ser excluyentes ya que los proveedores de contenido pueden no registrar con total claridad el contenido de su página o, incluso, no ser fehaciente.

2.2.1. RSAC

Con la idea original de proteger a los niños, ayudando a padres y profesores a controlar el acceso de éstos a Internet, surgió en 1994 el RSAC o *Recreational Software Advisory Council* que en 1999 se convertiría en el *ICRA* o *Internet Content Rating Association*. Esta iniciativa asociaba etiquetas de clasificación (meta-datos) a los contenidos de la red que son añadidas por los responsables del contenido de las páginas.

Las páginas web clasificadas con las etiquetas ICRA podrían ser leídas e interpretadas por el sistema de clasificación para poder aplicar el filtrado según la etiqueta correspondiente. Es el método más estandarizado.

Por ejemplo, el sitio web http://www.penthouse.com/ utiliza las etiquetas ICRA dentro de los meta datos de la cabecera tal y como se muestra a continuación:

```
<head>

   <link href="http://penthouse.com" rel="canonical"></link>
   <meta content="http://penthouse.com" http-equiv="Content-Location"></meta>
   <link href="http://penthouse.com?lang=italian" hreflang="it" rel="alternate"></link>
   <link href="http://penthouse.com?lang=swedish" hreflang="sv" rel="alternate"></link>
   <link href="http://penthouse.com" hreflang="en" rel="alternate"></link>
   <link href="http://penthouse.com?lang=french" hreflang="fr" rel="alternate"></link>
   <link href="http://penthouse.com?lang=chinese" hreflang="zh" rel="alternate"></link>
   <link href="http://penthouse.com?lang=japanese" hreflang="ja" rel="alternate"></link>
   <link href="http://penthouse.com?lang=portuguese" hreflang="pt" rel="alternate"></link>
   <link href="http://penthouse.com?lang=dutch" hreflang="nl" rel="alternate"></link>
   <link href="http://penthouse.com?lang=german" hreflang="de" rel="alternate"></link>
   <link href="http://graphics.penthouse.com/images/ph/css/header.css" type="text/css" rel="stylesheet"></link>
   <link href="http://graphics.penthouse.com/images/ph/favicon.ico" type="image/x-icon" rel="shortcut icon"></link>
   <link title="ICRA labels" type="application/rdf+xml" href="http://graphics.penthouse.com/images/ICRA_labels_rdf_adult.rdf" rel="meta"></link>
   <meta content="(pics-1.1 "http://www.icra.org/pics/vocabularyv03/"  l gen tr... 3 s 3 v 0 l 3 oa 0 ob 0 oc 0 od 0 oe 0 of 0 og 0 oh 0 c 3))" http-equiv="pics-Label"></meta>
   <meta content="RTA-5042-1996-1400-1577-RTA"  name="RATING"></meta>
   <meta content="text/html; charset=UTF-8" http-equiv="content-type"></meta>
   <link href="http://graphics.penthouse.com/css/live_cd/ph/spanish/0/flags-1309207811.css"  type="text/css" rel="stylesheet"></link>
   <link href="http://graphics.penthouse.com/css/live_cd/ph/spanish/0/fruit2-1418030718.css"  type="text/css" rel="stylesheet"></link>
   <link href="http://graphics.penthouse.com/css/live_cd/ph/spanish/0/global_ph-1430775255.css"  type="text/css" rel="stylesheet"></link>
   <title></title>
   <style></style>
   <style></style>
   <style type="text/css" media="screen"></style>

</head>
```

Ilustración 6. Ejemplo del etiquetado ICRA de una página web

```
    <meta content="(pics-1.1 "http://www.icra.org/pics/vocabularyv03/" l gen
tr… 3 s 3 v 0 l 3 oa 0 ob 0 oc 0 od 0 oe 0 of 0 og 0 oh 0 c 3))" http-
equiv="pics-Label"></meta>
```

En 2010 esta iniciativa se discontinuaría debido al poco uso por parte de los creadores de sitios web. Pero su trabajo no se perdería ya que el *FOSI* o *Family Online Safety Institute* tomaría el control del ICRA para continuar con su iniciativa desde un punto de vista educacional.

2.2.2. SafeSurf

Otra iniciativa basada en la idea de evitar que los niños accediesen a contenido para adultos a través de Internet fue SafeSurf. Nació en 1995 y se basaba en la misma filosofía de etiquetar sitios.

Como diferencia con el anterior, SafeSurf cuenta con más categorías de clasificación y diferencia niveles ya que no sólo describe objetivamente el contenido de una página sino también cómo se presenta ese contenido. Por ejemplo, no es lo mismo hablar de *'desnudos'* (categoría "nudity") en un texto médico que en una página dedicada a contenidos eróticos. Así se ofrece una mayor flexibilidad en el control de los contenidos.

Un ejemplo del etiquetado que utiliza SafeSurf sería:

```
<META http-equiv="PICS-Label" content='(PICS 1.0
"http://www.classify.org/safesurf/" l r (SS~~000 4 SS~~001 5 SS~~004 2
SS~~007 2 SS~~008 3))'>
```

2.3. Análisis de textos

Pero cuando el contenido o no ha sido revisado o no viene marcado, hay que implementar otra serie de análisis.

Cualquier sistema de filtrado de contenidos se basa en el análisis de tráfico generado para transferir desde el servidor que lo alberga, el contenido solicitado por un usuario. En el caso del protocolo HTTP, la comunicación entre el navegador y el servidor se lleva a cabo en dos etapas:

- El navegador del usuario realiza una *solicitud HTTP*

```
GET http://www.elmundo.es/ HTTP/1.1
Host: www.elmundo.es
Proxy-Connection: keep-alive
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Upgrade-Insecure-Requests: 1
User-Agent: Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.155 Safari/537.36
DNT: 1
Accept-Encoding: gzip, deflate, sdch
Accept-Language: es-ES,es;q=0.8,en;q=0.6
```

- El servidor procesa la solicitud y envía una *respuesta HTTP*

```
HTTP/1.1 200 OK
Server: nginx/1.4.4
Date: Fri, 14 Aug 2015 09:03:19 GMT
Content-Type: text/html
Vary: Accept-Encoding
Vary: User-Agent
Set-Cookie: ELMUNDO_idusr=Vc2u18CoFZAAACBUMSA-348d8181a4111ace1687fa72fdefd46a; expires=Mon, 13-Aug-2018 09:03:19 GMT; path=/; domain=.elmundo.es
Set-Cookie: ELMUNDO_pref=%7B%22v%22%3A%22n%22%2C%22d%22%3A%22e%22%2C%22u%22%3A0%2C%22c%22%3A%222%22%7D; expires=Mon, 13-Aug-2018 09:03:19 GMT; path=/; domain=.elmundo.es
Connection: Keep-Alive
Keep-Alive: timeout=60, max=8

12e0
<!DOCTYPE html>
```

```
<html lang="es">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-15"/>
<…>
</head>
<body>
<…>
</body>
</html>
```

En cada una de estas etapas podremos realizar diferentes técnicas de análisis como veremos a continuación.

2.3.1. Análisis de la URL

Cuando el navegador realizar una solicitud HTTP, envía un comando (GET) con información sobre el contenido que está solicitado (la URL).

Utilizando los diccionarios se podrían realizar detecciones a partir de la URL contenida en el comando GET. Así se podría realizar un análisis de URL como el detallado a continuación:

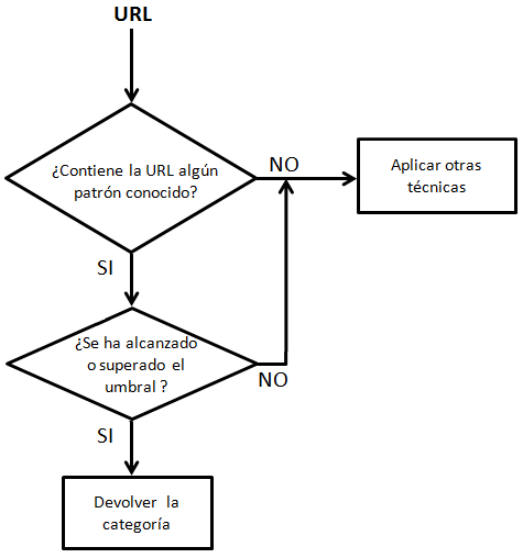


Ilustración 7. Diagrama del análisis de URL

2.3.2. Análisis del contenido devuelto por el servidor

Pero cuando la URL no presenta ningún signo de peligro, otro nivel de detección sería realizar un análisis del propio contenido que devuelve el servidor que aloja el recurso solicitado por el usuario.

Este análisis se llevaría a cabo sobre la respuesta que enviase el servidor y que incluye el contenido del recurso solicitado por un usuario. El análisis de contenido seguiría un flujo semejante al análisis de URL que acabamos de ver:

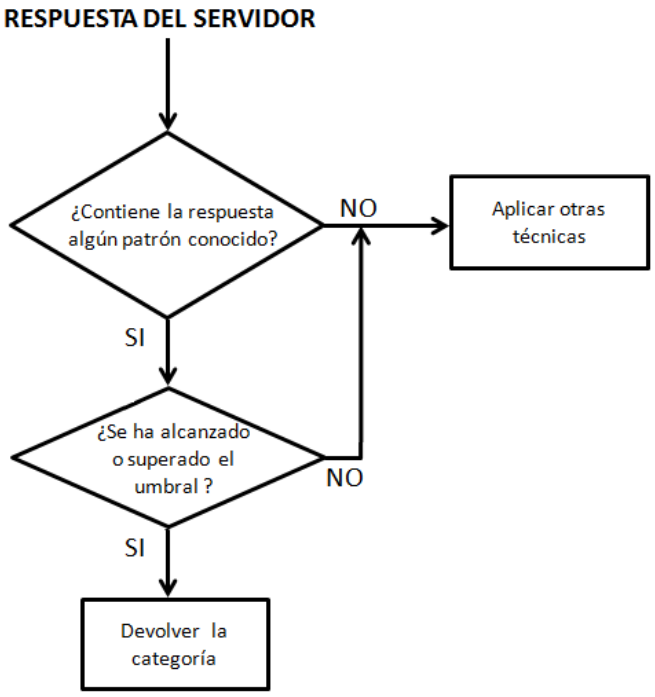


Ilustración 8. Diagrama del análisis de un contenido devuelto

3. Arquitectura de un sistema de filtrado de contenidos

Comencemos a darle forma al sistema de filtrado de contenidos basados en las diferentes técnicas vistas anteriormente.

3.1. Diseño de la solución

La arquitectura general del sistema de filtrado de contenidos será la siguiente:

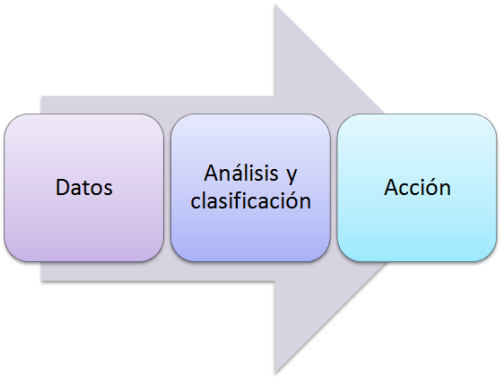


Ilustración 9. Diagrama básico de un sistema de análisis de contenidos

A continuación se muestra el diagrama anterior complementado con las técnicas de análisis vistas anteriormente y que se utilizarán en este proyecto:

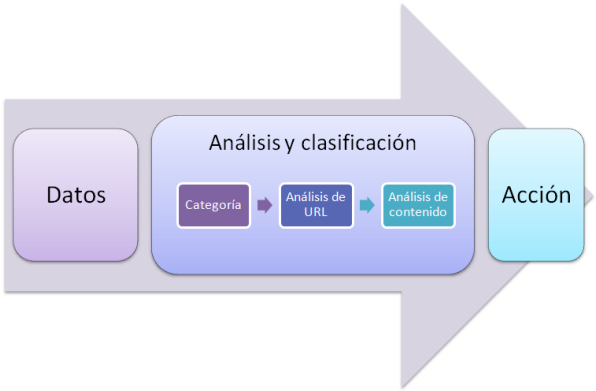


Ilustración 10. Diagrama general de un sistema de análisis de contenidos

3.2. Descripción e implementación

Partiremos de la solicitud de acceso a una página web como el dato inicial que recibirá nuestro sistema de análisis de contenidos. Éste deberá analizar y clasificar dicha página para que de acuerdo a la configuración del sistema, se pueda acceder o bloquear la solicitud.

Desde el punto de vista de la arquitectura de red, el sistema diseñado debería contar con las siguientes características:

- *Flexibilidad.* La solución se debería poder adaptar a las necesidades particulares de las arquitecturas o sistemas existentes donde se integre dicha solución.
- *Escalabillidad.* Para permitir planificar la inversión total necesaria de forma secuencial integrando la solución según las necesidades del sistema en el que se integre sin perder calidad del servicio ofrecido.
- *Facilidad de ampliación.* El sistema se diseñará desde sus inicios para que pueda ir creciendo paulatinamente con total facilidad según las futuras necesidades.
- *Eficiencia.* Al diseñar el sistema de forma flexible y escalable, es sencillo implementar soluciones a medida según las necesidades de los diferentes clientes manteniendo la máxima eficiencia.

4. Constitución de un sistema modular de filtrado de contenidos

En este apartado veremos qué elementos son necesarios para poder completar una solución de filtrado de contenidos.

4.1. Descripción del proceso de análisis de contenidos

En el siguiente diagrama se puede observar un esquema del proceso completo de la solución de filtrado de contenidos de una página web.

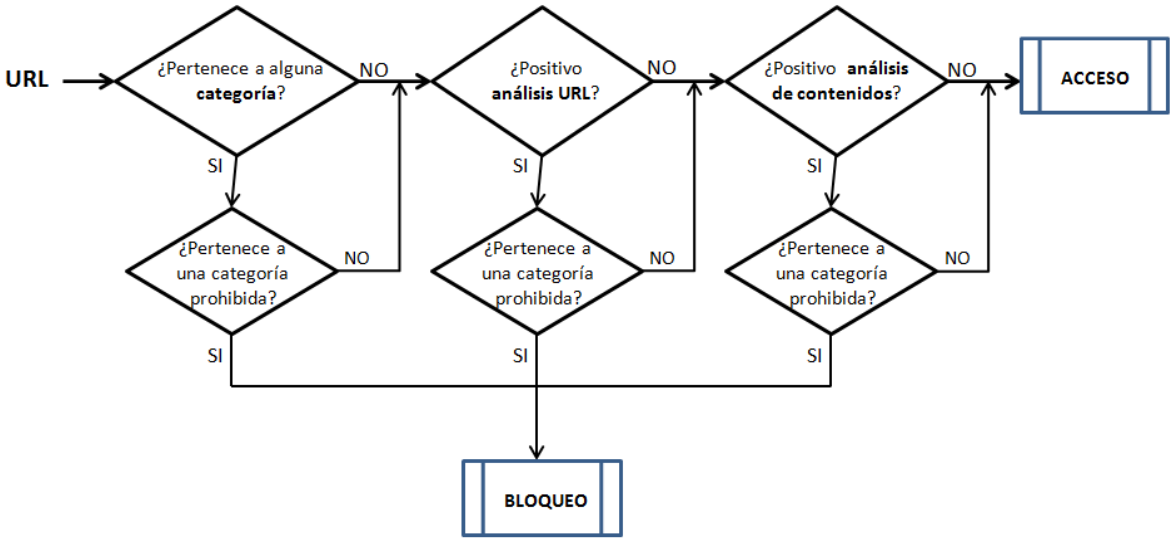


Ilustración 11. Diagrama completo del sistema de análisis de filtrado

4.2. Página de bloqueo

Cuando la petición de la página web no se bloquea, el usuario accede al recurso solicitado. Pero en el caso que se deniegue su petición, se debe servir una página Web que sustituya a la solicitada por el usuario. Esta página web debe informar que el contenido solicitado ha sido bloqueado por contener contenido inadecuado.

A continuación se muestra un ejemplo sencillo de la página que podría visualizar un usuario en el caso de que su petición fuese bloqueada.

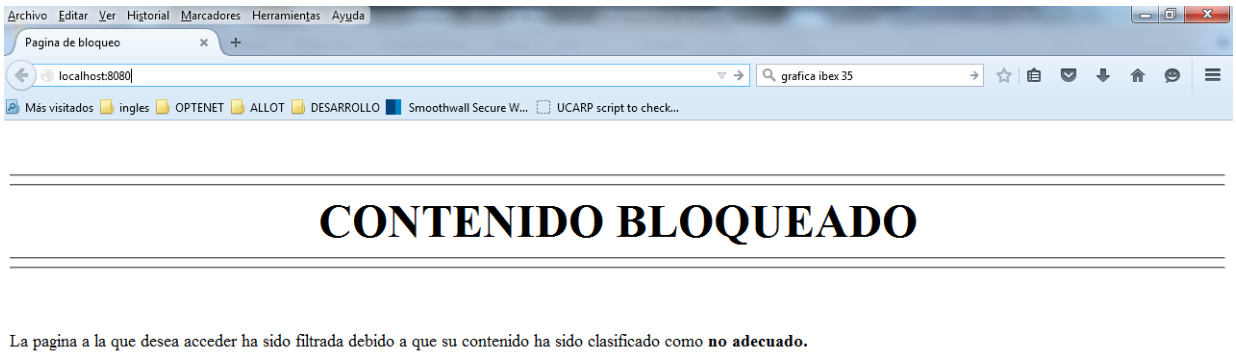


Ilustración 12. Ejemplo de página de bloqueo

4.3. Experiencia de usuario

A continuación veremos la experiencia de usuario que un usuario percibirá ante una solución de filtrado de contenidos como la anteriormente descrita.

Imaginemos que un usuario quiere acceder a la web de www.playboy.com. Esta página web está categorizada como *pornografía* luego dicho contenido estaría bloqueado. En cuanto la petición fuese recibida por el sistema de filtrado, sería bloqueada mostrando la página de bloqueo al usuario sin que su solicitud llegase a Internet.

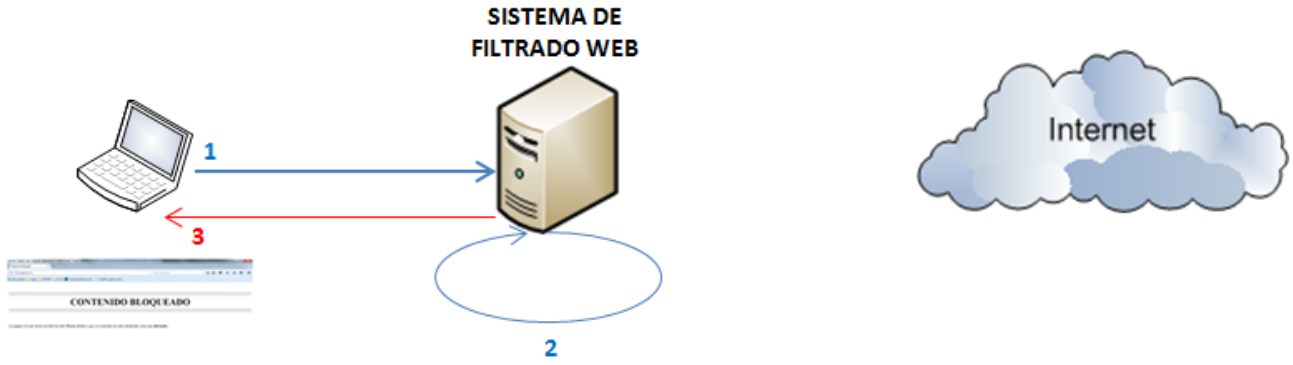


Ilustración 13. Experiencia de usuario de un bloqueo por categoría

Número	Descripción
1	El usuario solicita la página www.playboy.com
2	El sistema de filtrado da positivo por categoría o análisis de URL.
3	El sistema de filtrado presenta una página de bloqueo al usuario.

Tabla 1: Flujo de comunicación de un bloqueo por categoría

El flujo anterior también aplicaría en el caso de que el análisis de URL diese positivo.

Sin embargo, en el caso que el usuario intentase acceder a una página web no categorizada y que el análisis de URL no diera positivo, habría que analizar el contenido de la página para decidir si se le sirve o no al usuario. En ese caso, la petición de la página web lanzada por el usuario llegaría a Internet.

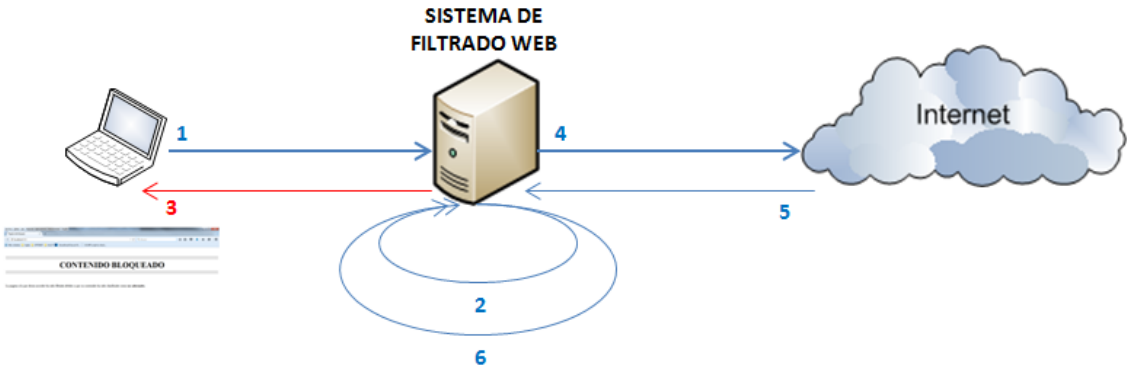


Ilustración 14. Experiencia de usuario de un bloqueo por análisis de contenido

Número	Descripción
1	El usuario solicita una página Web.
2	El sistema de filtrado no da positivo por categoría o análisis de URL.
4	El sistema de filtrado solicita la página a Internet.
5	El sistema de filtrado recibe la página solicitada por el usuario.
6	El sistema de filtrado analiza el contenido de la página recibida y da positivo.
3	El sistema de filtrado presenta una página de bloqueo al usuario.

Tabla 2: Flujo de comunicación de un bloqueo por análisis de contenido

Por último, cuando un usuario intenta acceder a una página web de contenido lícito, aunque se realizan todos los procesos de análisis, el usuario no percibe que se realice ningún tipo de análisis y accede a la página web solicitada (en el ejemplo, www.elmundo.es).

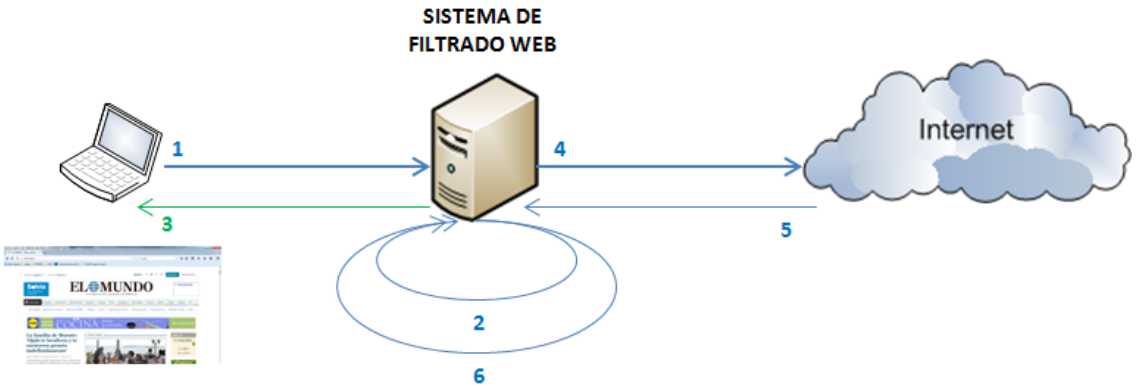


Ilustración 15. Experiencia de usuario de un acceso al contenido web

Número	Descripción
1	El usuario solicita la página web www.elmundo.es
2	El sistema de filtrado no da positivo por categoría o análisis de URL.
4	El sistema de filtrado solicita la página a Internet.
5	El sistema de filtrado recibe la página solicitada por el usuario.
6	El sistema de filtrado analiza el contenido de la página recibida y da positivo.
3	El sistema de filtrado devuelve la página solicitada por el usuario.

Tabla 3: Flujo de comunicación de un acceso al contenido web

4.4. Diagrama de comunicación

El protocolo a través del cual se realizarán las peticiones de contenidos web y sobre el que centraremos nuestra atención será el protocolo HTTP. Únicamente nos centraremos en la parte de la comunicación HTTP en la que se realiza el filtrado de contenidos que es la base de nuestro estudio.

La resolución DNS y la apertura/cierre de la conexión TCP/IP serán transparentes en el análisis de la solución de filtrado.

A continuación veremos cómo interaccionan los diferentes elementos implicados en la comunicación para los distintos casos de uso vistos en el apartado anterior durante la petición o transmisión de los datos.

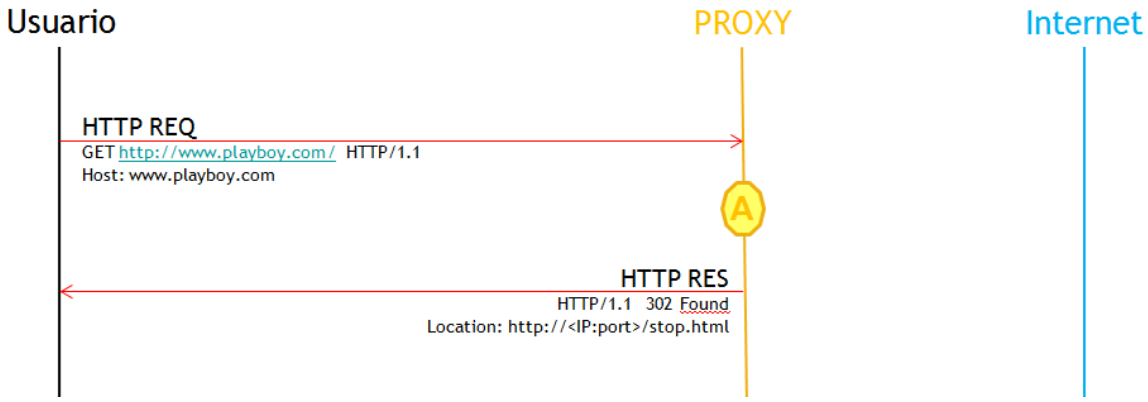


Ilustración 16. Diagrama de comunicación en un bloqueo por categoría/ análisis de url

El intervalo 'A' representa el momento en el que la página web solicitada es analizada y el resultado de la detección es positivo (ya sea por categoría o análisis de URL).

El usuario no accede al contenido solicitado.

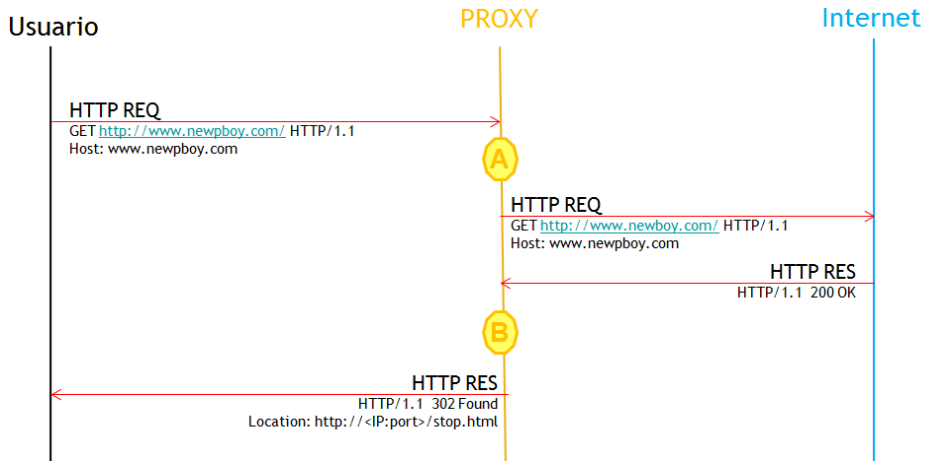


Ilustración 17. Diagrama de comunicación en un bloqueo por análisis de contenido

El intervalo 'A' representa la misma acción que en el diagrama anterior. En este caso la detección es negativa tanto por categoría como por análisis de URL. Se prosigue con la petición al servidor de la página web solicitada por el usuario.

El intervalo 'B' representa el análisis del contenido de la página web devuelta por el servidor que aloja dicho sitio. En este caso, la detección resulta positiva. El usuario no accede al contenido solicitado.

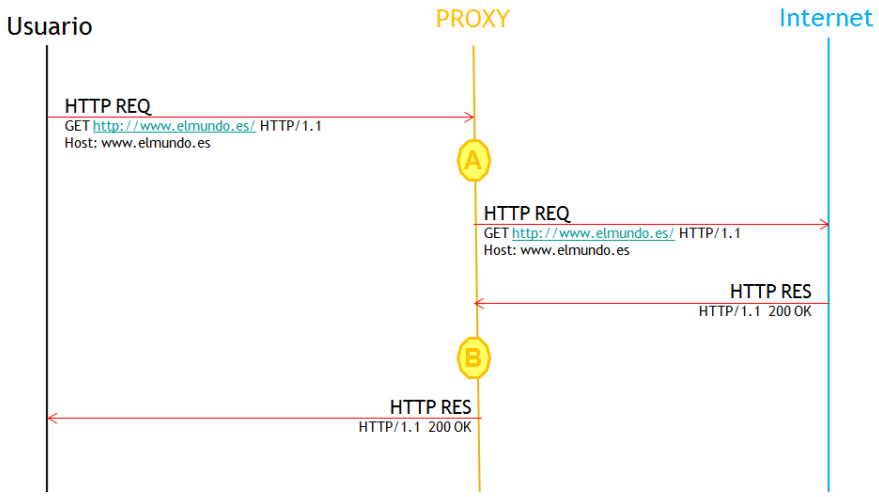


Ilustración 18. Diagrama de comunicación de un acceso al contenido web

El significado de los intervalos 'A' y 'B' coinciden con la descripción del diagrama anterior. Sin embargo, en este caso el resultado de la detección es negativo. El usuario accede al contenido solicitado.

CAPÍTULO 3: Entorno de pruebas y resultados obtenidos

1. Introducción

Una vez expuesto el fundamento técnico del sistema de filtrado de contenidos web, en este capítulo se describe el entorno de pruebas que se va a utilizar y los experimentos realizados y resultados obtenidos.

2. Entorno de pruebas

A continuación veremos una descripción de la maqueta de pruebas desplegada y utilizada para este proyecto.

2.1. Características del servidor proxy

Los requisitos mínimos sobre los que se debe instalar el software que implementa la solución de filtrado son los siguientes:

- 1 interfaz de red (compartida para el plano de gestión y el de datos)
- 4GB RAM
- 20GB HDD

La solución de filtrado se ha desplegado en un entorno virtual VMWare ESXi 4 simulando un servidor físico que cumpliese los requisitos mínimos anteriores.

2.1.1. Procesador

Se utilizará un procesador Intel(R) Xeon(R) CPU E3-1245 V2 @ 3.40GHz con cuatro cores.

2.1.2. Tamaño de disco

El tamaño del disco asignado para la instalación de la solución será de 22GB. Tras la instalación, se ocuparán 7.4GB dejando el resto para la escritura de datos necesarios, por ejemplo, para el logado de las acciones realizadas por la solución.

2.1.3. Memoria RAM

La memoria RAM disponible será de 6113124 kB.

2.1.4. Interfaces de red

Una única interfaz sera configurada para recibir las peticiones de los usuarios y acceder a Internet. Por motivos de seguridad, para el acceso a dicho entorno se utilizará en la configuración de red, direccionamiento privado con acceso público mediante NAPT.

2.1.5. Sistema Operativo

El sistema operativo sobre el que corra la solución de filtrado será una distribución GNU/Linux de 64 bits.

2.2. Arquitectura de red

A continuación se detallará el diagrama de red utilizado y se describirán cómo se han implementado cada uno de los elementos que conforman el entorno de pruebas.

2.2.1. Topología de red

En redes de comunicación las soluciones se suelen desplegar utilizando topologías de red de tipo transparente o proxy.

Las diferencias entre estos modos de despliegue no sólo aplican a nivel de red, existen también otras:

- Transparente o bridge – todo el tráfico que genera un usuario atraviesa el sistema de filtrado. Luego esta topología de red no modifica la identidad entre los extremos ni a nivel de direccionamiento IP ni de MAC.

- Proxy – sólo el tráfico proxificado es analizado. En este modo de despliegue, se modifican los paquetes en cuanto a nivel de direccionamiento IP y MAC.

El empleo de una u otra topología vendrá fuertemente determinado por las funcionalidades del servicio solicitado por un cliente.

En entornos empresariales, los usuarios suelen iniciar sesión dentro de dominios definidos previamente por los administradores de la empresa. Así se controla el acceso a los recursos de la misma según los privilegios asignados a cada grupo o usuario. En soluciones de filtrado, se suele utilizar esta característica implementándose la autenticación de usuario de tipo básico o contra bases de datos externas utilizando protocolos como NTLM o Kerberos. En estos casos, un despliegue en transparente no tendría sentido.

Sin embargo, para poder utilizar un proxy es necesario que cada dispositivo que utilice un usuario para acceder a Internet sea configurado para que el tráfico generado por él lo reciba del proxy. En este caso, si es necesario que el servicio se ponga en marcha sin que el usuario sea consciente o no tenga que realizar ningún cambio en su configuración, habría que desplegar en modo transparente y la autenticación habría que salvaguardarla de otra manera, siempre que fuese necesaria.

Las soluciones son muy variadas dependiendo de los requisitos. Por tanto, en este proyecto implementaremos para nuestras pruebas un despliegue en modo proxy explícito debido a la sencillez con la que se puede generar tráfico simulando a varios usuarios accediendo de forma simultánea o el uso de diversos dispositivos sin tener que realizar apenas modificaciones a nivel de red.

2.2.2. Diagrama de la arquitectura de red

A continuación se muestra un diagrama con la arquitectura de red utilizada para este entorno de pruebas.

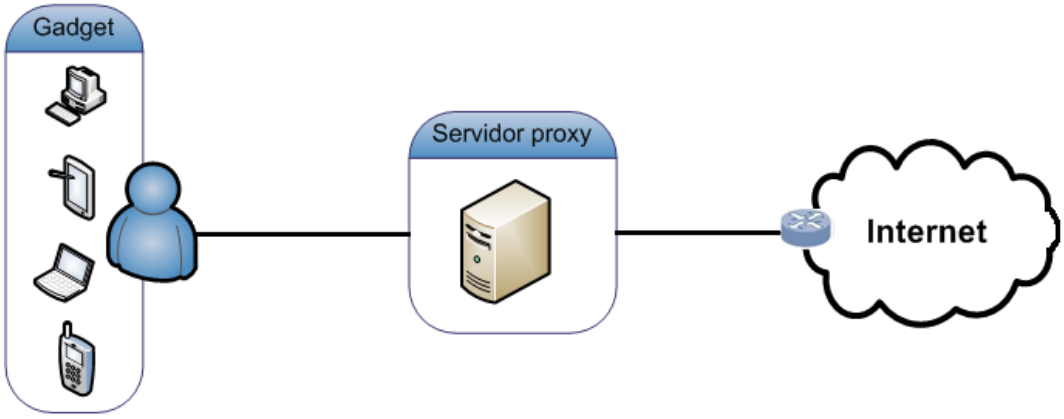


Ilustración 19. Arquitectura de red del entorno de pruebas

Como ya hemos dicho, en un entorno real, para poder utilizar el anterior esquema el tráfico generado por el usuario deberá ser dirigido hacia el proxy. Esto es, desde cualquier dispositivo, será necesario modificar la configuración del navegador utilizado para que las peticiones realizadas por el usuario sean enviadas explícitamente al servidor proxy y éste pueda actuar como sistema de filtrado de contenidos web.

A continuación se describen cada uno de los elementos que aparecen en ilustración anterior. En el diseño de los mismos se ha pensado en facilitar la medida y realización de las pruebas.

2.2.2.1. Internet

La *nube de Internet* representa las páginas web disponibles en todos los servidores por todo el mundo. Para controlar el entorno, esta nube será simulada mediante una estructura de archivos alojada en un servidor web interno. Se usará un servidor Apache y los contenidos estarán accesibles desde `/var/www/html`.

Los contenidos albergados no serán reales sino generados artificialmente para que contengan unos ficheros con una serie de características conocidas previamente tales como el tamaño, contenido y tipo de contenido (accesible/bloqueado).

Se definirán dos directorios según la naturaleza del contenido web (*filter*: bloquear, *text*: acceder y *notex*t: imágenes) y dentro de ellos se alojarán los recursos según su tamaño:

```
/<test_name>/filter/500
                   /1500
                   /2500
                   /3500
                   /5000
                   /50000
                   /100000
                   /200000
/<test_name>/text/500
                  /1500
                  /2500
                  /3500
                  /5000
                  /50000
                  /100000
                  /200000
/<test_name>/notext/500
                   /1500
                   /2500
                   /3500
                   /5000
                   /50000
                   /100000
                   /200000
```

El formato de las páginas web de prueba es .html y éstas no estarán necesariamente bien formadas:

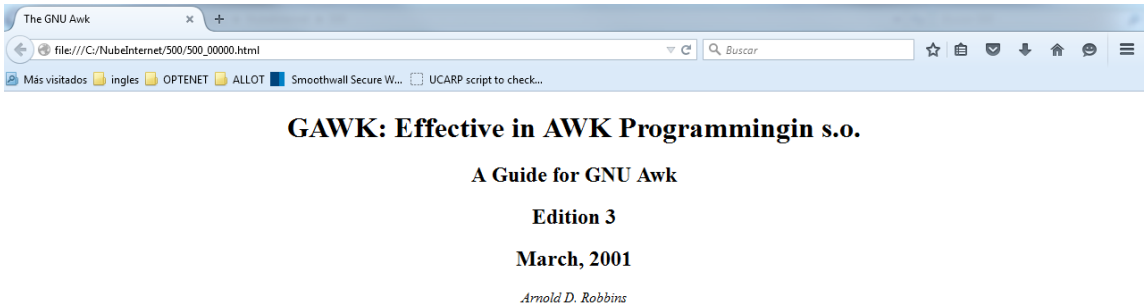


Ilustración 20. Ejemplo de página web de prueba accesible de 500 bytes

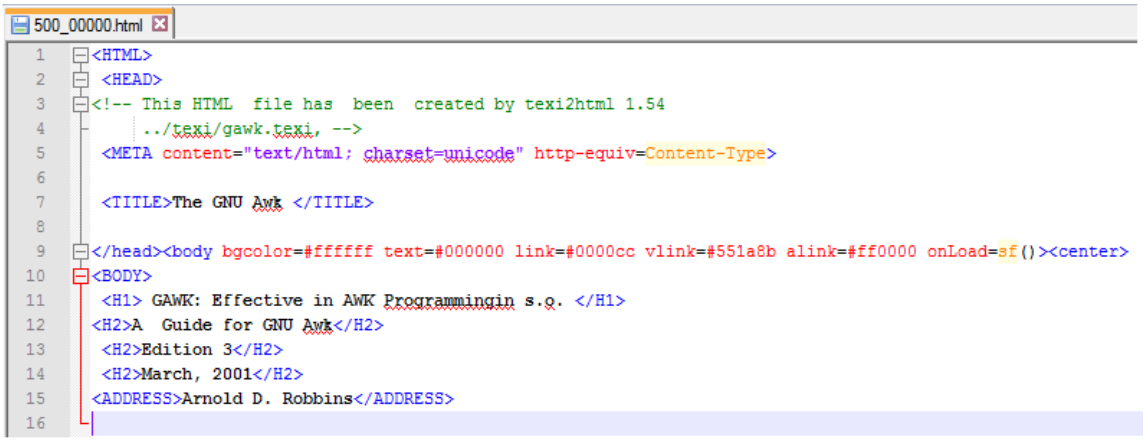


Ilustración 21. Código HTML del contenido anterior

2.2.2.2. Cliente PC

Desde un ordenador se simularán las peticiones http necesarias (tanto las simultáneas como las secuenciales) para realizar las pruebas. El formato de las peticiones a los contenidos alojados en el servidor Apache será el siguiente:

```
http://<server_IP>/<test_name>/<file_type>/<file_size>/<file_name>
```

donde `<file_size>` representa el tamaño del recurso solicitado en bytes.

Por ejemplo:

```
http://172.16.0.50/latencia/text/500/500_00000.html
```

Todas las conexiones utilizarán el protocolo HTTP/1.1

En algunas pruebas, se utilizará el comando de Linux *curl* para simular las peticiones que un usuario:

```
curl http://172.16.0.50/latencia/text/500/500_00000.html
```

2.2.2.3. Servidor proxy

El detalle de este elemento se ha descrito anteriormente en el apartado "2.1 Características del servido proxy".

2.3. Pruebas y resultados obtenidos

Una vez que ya tenemos definido el entorno de pruebas semejante al que nos podríamos encontrar en un cliente real y todos los elementos necesarios, procederemos a realizar una serie de pruebas para determinar:

- El retardo que introduce la solución de filtrado en la comunicación
- El momento en el que la solución de filtrado deja de responder de forma óptima según la cantidad de tráfico recibido
- Medidas de falsos positivos y negativos

2.3.1. Modelado del tráfico

El tipo de peticiones utilizadas afectará en mayor medida cuando evaluemos el rendimiento del sistema. Para generar las peticiones web de forma simultánea se va a emplear una herramienta desarrollada por un equipo externo. Algunos parámetros útiles que se le pueden pasar por parámetro a esta herramienta son:

- La lista de URLs que se quieren solicitar
- La dirección IP y puerto del proxy
- Número máximo de peticiones
- Número máximo de peticiones por segundo
- Máximo ancho de banda (Kbps)
- Tiempo de espera para que terminen todas las peticiones (seg)
- El fichero en el que volcar los datos de salida

Como resultado, proporcionará cierta información estadística como la mostrada a continuación:

- Número de peticiones lanzadas
- Número de peticiones lanzadas por segundo
- Ancho de banda consumido (Kbps)
- Tiempo medio en el que se ha recibido la respuesta a una petición (medido en segundos).
- Número de peticiones completadas según el código de respuesta:
  - *200 OK* > contenido accedido
  - *302 Found* > contenido bloqueado (indica redirección a la página de bloqueo)
- Número total de bytes recibidos desde el comienzo de la simulación

Un ejemplo de los valores devueltos a la salida de la ejecución de esta herramienta son:

```
15/02/15 13:13:46
Peticiones lanzadas  : 29878
Peticiones codigo == 200: 28279
Peticiones codigo == 302: 656
Tiempo empleado      : 156.781 sg
Bytes recibidos      : 44933105
Ancho Banda consumido: 2798.799 Kbps
Peticiones /sg       : 190.6
Tiempo medio en servir una peticion    : 0.859984 sg
Repartidas 29889 Terminadas 29878
```

El tipo de contenidos solicitados también afectará a los resultados.

Actualmente la mayoría de páginas de Internet contienen, además de textos, otra serie de contenidos como imágenes, vídeos, enlaces a otras páginas u hojas estilo. Por tanto, se deberán incluir contenidos que simulen este tipo de solicitudes (esto es, los recursos o páginas alojados dentro del directorio *notex*t de la nube de Internet).

Con respecto a los bloqueos, en general, no suelen ser predominantes sino todo lo contrario, minoritarios.

Las peticiones http seguirán el siguiente esquema:

- Accesos mayoritarios a contenidos no textuales
- Bajo número de bloqueos

Para obtener el listado de URLs con el que vamos a trabajar, utilizaremos una herramienta propietaria desarrollada por un equipo externo.

El fichero de URLs se definirá a partir de los siguientes parámetros:

- Tamaño medio de una petición (bytes)
- Porcentaje de contenidos con texto (incluirá URLs *text y filter*)
- Porcentaje de contenidos bloqueadas

Tomaremos un tamaño medio por petición de 35kB, un 30% de URLs con texto (el 70% restante, contendrá otro tipo de contenido o URLs alojadas en *notext*) y un 20% de contenidos bloqueados.

Para las pruebas de rendimiento se generará un número definido de peticiones HTTP. El patrón de tráfico enviado se puede representar mediante una serie de pulsos como se muestra a continuación:

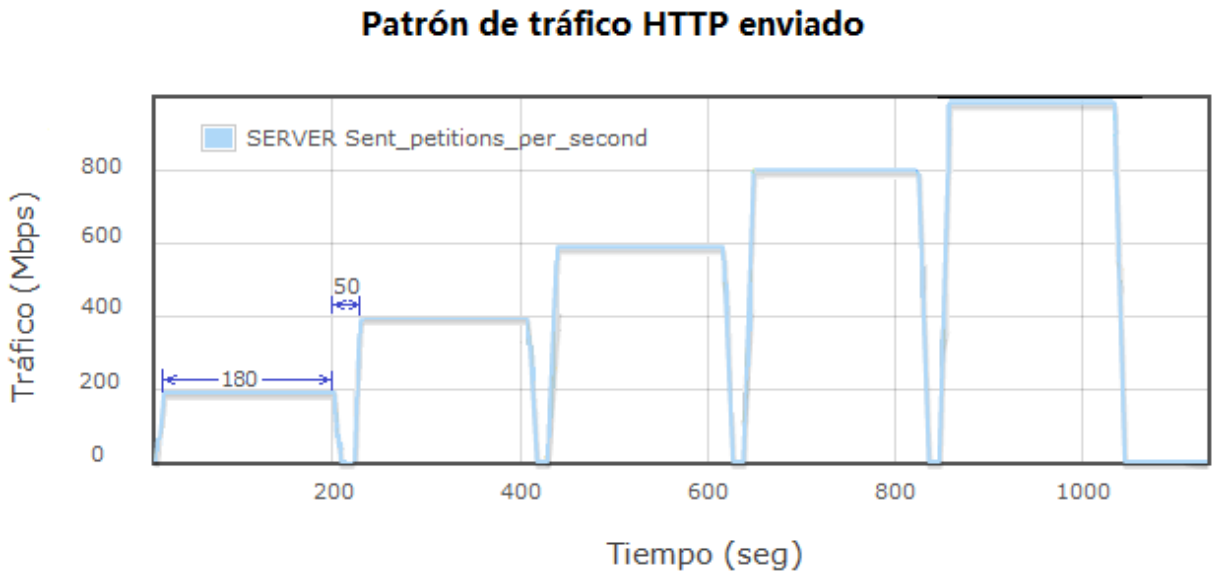


Ilustración 22. Representación del tráfico enviado

Las peticiones por segundo (pps) necesarias para modelar un tráfico como el anterior se muestran en la siguiente tabla:

Peticiones por segundo (pps)	Ancho de banda equivalente (Mbps)
700	200
1395	400
2090	600
2790	800
3485	1000

Tabla 4: pps - Mbps

Es importante señalar que los resultados que se van a obtener en las pruebas son estimaciones y que variarán dependiendo de la naturaleza del tráfico utilizada.

2.3.2. Latencia

Al introducir un elemento nuevo entre la comunicación directa del usuario y el servidor que aloja el contenido web, cabe esperar que este nuevo sistema introduzca algún tipo de retardo.

Esta latencia o retardo será introducida por el procesamiento interno que realiza la solución de filtrado sobre el recurso solicitado a Internet, aunque el usuario no será consciente de ello.

En esta primera prueba mediremos el tiempo extra que el sistema de filtrado introduce en la comunicación. En este caso, los contenidos solicitados serán páginas reales de Internet. Se utilizará un script que lanzará de forma simultánea y controlada las peticiones simultáneas a Internet. Para medir el tiempo de respuesta, se utilizarán los comandos de Linux *time curl* y su valor devuelto *real*.

```
[root@testbed-00-0C-29-82-4B-DD database]# time curl www.google.es
 ***** ***** ***** ***** ***** *****

real    0m0.082s
user    0m0.001s
sys     0m0.002s
```

En la siguiente gráfica se representa la diferencia en los tiempos de respuesta medidos al realizar las peticiones a través de la solución de filtrado y sin pasar por ella.

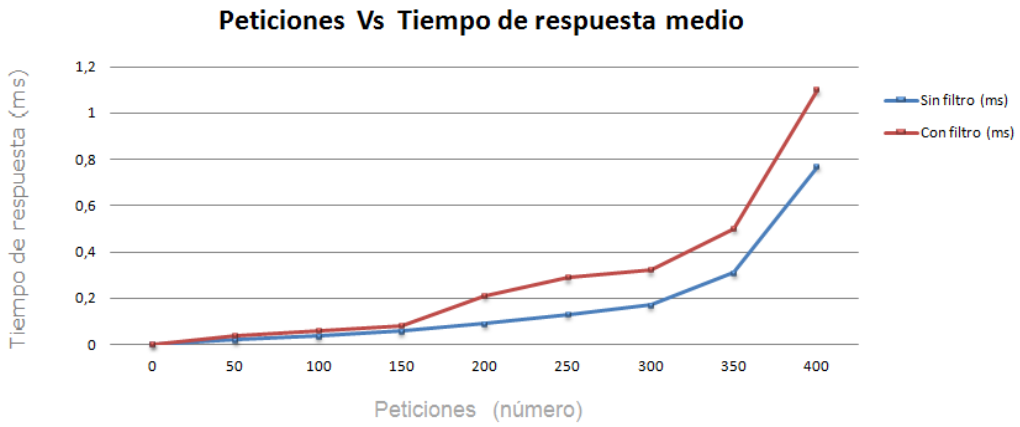


Ilustración 23. Peticiones Vs Tiempo de respuesta medio

En el siguiente diagrama de barras se representa la diferencia en los tiempos medidos anteriormente para una mejor medición:

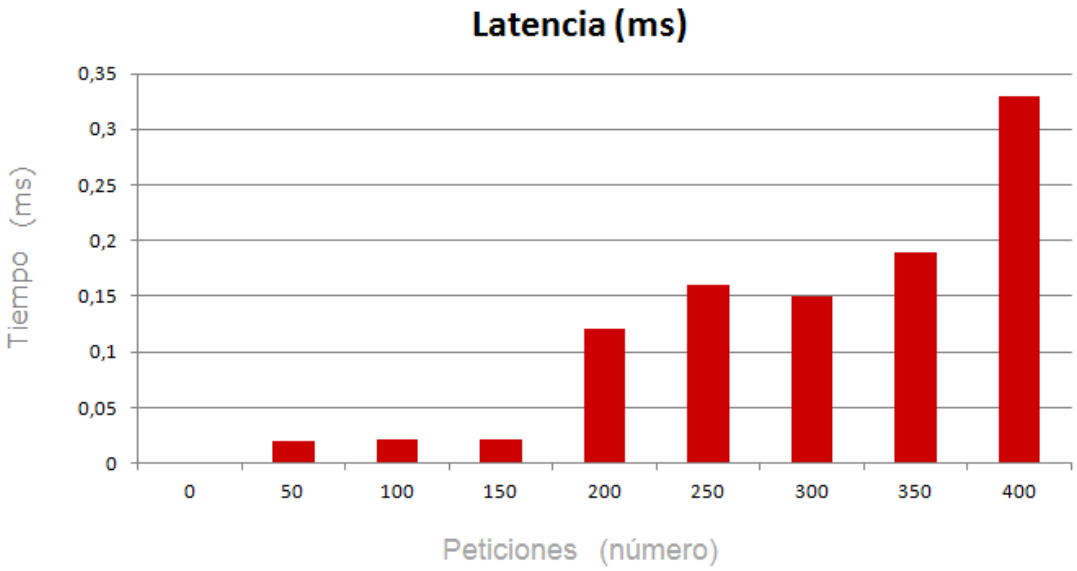


Ilustración 24. Diferencia de latencia medidas con y sin filtro

A la vista de esta última gráfica podemos ver que la latencia introducida por el sistema de filtrado es inferior a 0.4 ms.

2.3.3. Rendimiento

La interfaz de servicio configurada en la máquina virtual donde se ha desplegado la solución de filtrado de contenidos es de 1 Gb. Por tanto, este será el ancho de banda o valor máximo de datos que podrá recibir la plataforma proxy.

En estas pruebas mediremos la respuesta del sistema ante incrementos de tráfico y su valor máximo de funcionamiento sin degradación del servicio.

En el siguiente gráfico se muestra cómo se comporta el sistema en cuanto a la capacidad de procesamiento de peticiones del sistema del filtrado.

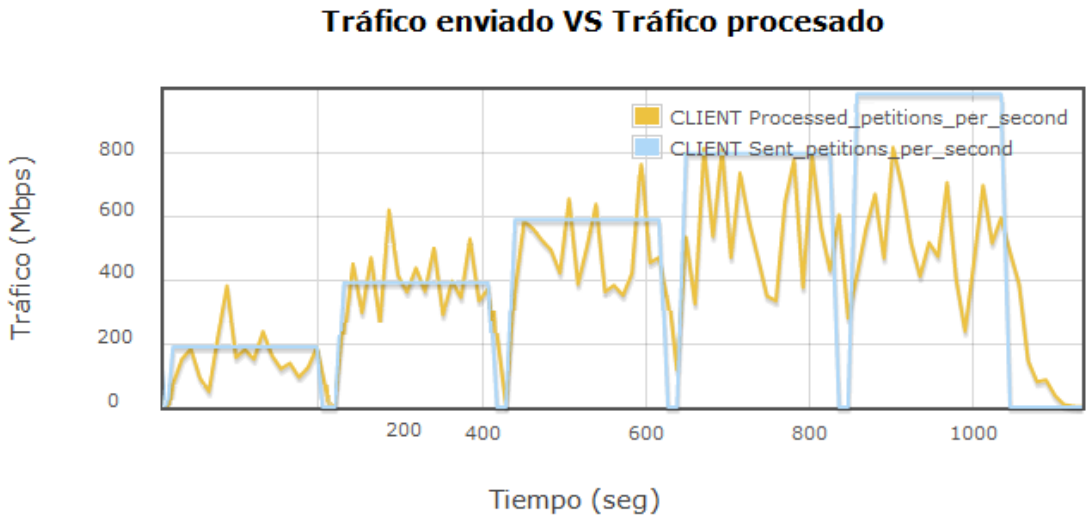


Ilustración 25. Tráfico enviado VS Tráfico procesado

Del anterior gráfico, vemos que cuando se envían aproximadamente unas 2090 pps (600 Mbps) el sistema de filtrado no es capaz de procesar todas las peticiones que le llegan. Por tanto, su límite se encontrará en el intervalo de 400-600 Mbps.

Sobre la gráfica anterior, podríamos aproximar que el valor máximo de procesamiento de la solución de filtrado en modo proxy sería de unos 480 Mbps.

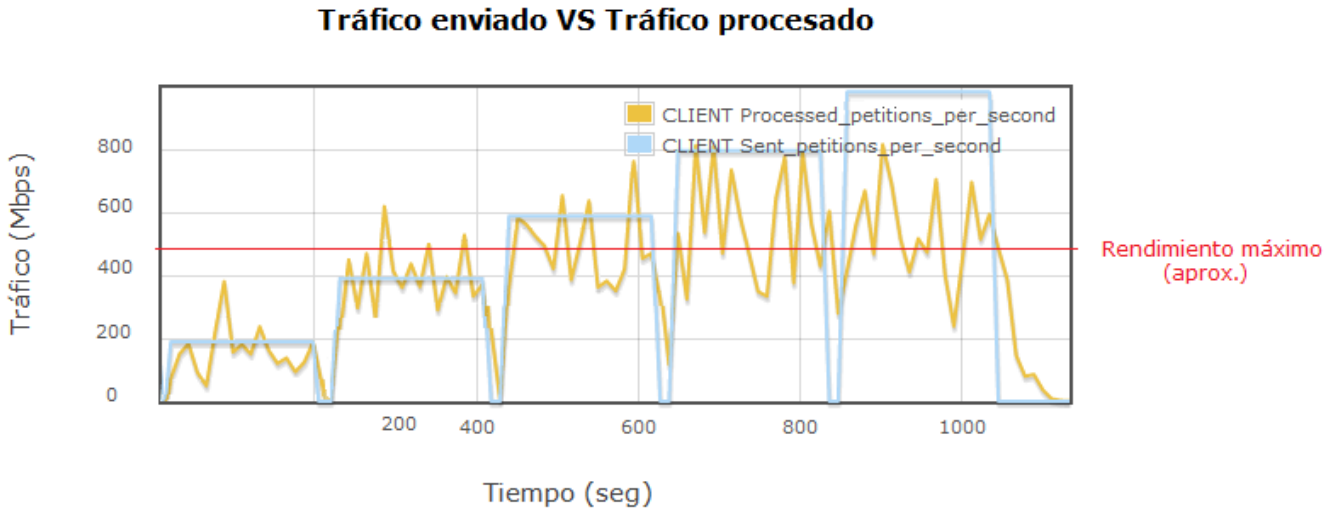


Ilustración 26. Estimación del rendimiento

Un rendimiento de este valor es pésimo al representar un rendimiento inferior al 50% de la capacidad de la conexión. La razón es que al logar información sobre las peticiones procesadas, estas escrituras son muy lentas debido al entorno virtual (y la simulación de los discos físicos mediante discos virtuales). Si se deshabilitan los logs, el rendimiento mejora pero no tendría sentido evaluar una plataforma sin ellos ya que no podrían ofrecer información sobre las acciones realizadas y un análisis de incidencias no sería posible.

Por tanto, se concluye que existe una gran limitación al usar una máquina virtual en vez de un servidor real.

2.3.4. Falsos positivos/negativos

Para evaluar la eficiencia del análisis que realiza la solución de filtrado vamos a utilizar el sitio web Alexa (www.*alexa.com)* que provee información sobre las páginas más visitadas atendiendo a diferentes criterios.

En cada apartado se han realizado una serie de gráficas donde se representarán los resultados de la categorización que la solución de filtrado ha realizado sobre cada una de las listas de URLs tomadas de Alexa. Los sitios han sido revisados manualmente para verificar la categorización realizada.

De la información disponible en esta web, para el estudio de falsos positivos y negativos se han evaluado las siguientes listas:

- Los 100 sitios de Internet más visitados mundialmente
- De los 500 sitios de Internet más visitados categorizados según Alexa de contenido adulto:
    - Los 100 primeros sitios o "Top 100 de contenidos para adultos"
    - Los sitios situados entre las posiciones 400 y 500 de contenidos para adultos.

2.3.4.1. Top 100 mundial

El resultado del análisis de contenidos para los 100 sitios de Internet más visitados mundialmente es:

Ilustración 27. Análisis de contenidos de los 100 sitios más visitados

A la vista de la gráfica anterior, se registró un único sitio que no estaba clasificado (*Googleadservices.com*) y no se encontró ningún falso positivo.

De acuerdo a Alexa, los sitios más visitados corresponderían a páginas categorizadas como buscadores (*google.com*) y de compras on-line (*amazon.com*).

Los sitios categorizados como 'Otros' pertenecen a temáticas como viajes, empleos, foros o material de referencia.

2.3.4.2. Top de contenidos de adultos

Para focalizar la detección de errores en la clasificación, analizaremos ahora los resultados para los 100 sitios más visitados de contenido de adultos. Entre los que a priori, esperamos que destaque la categoría de pornografía.

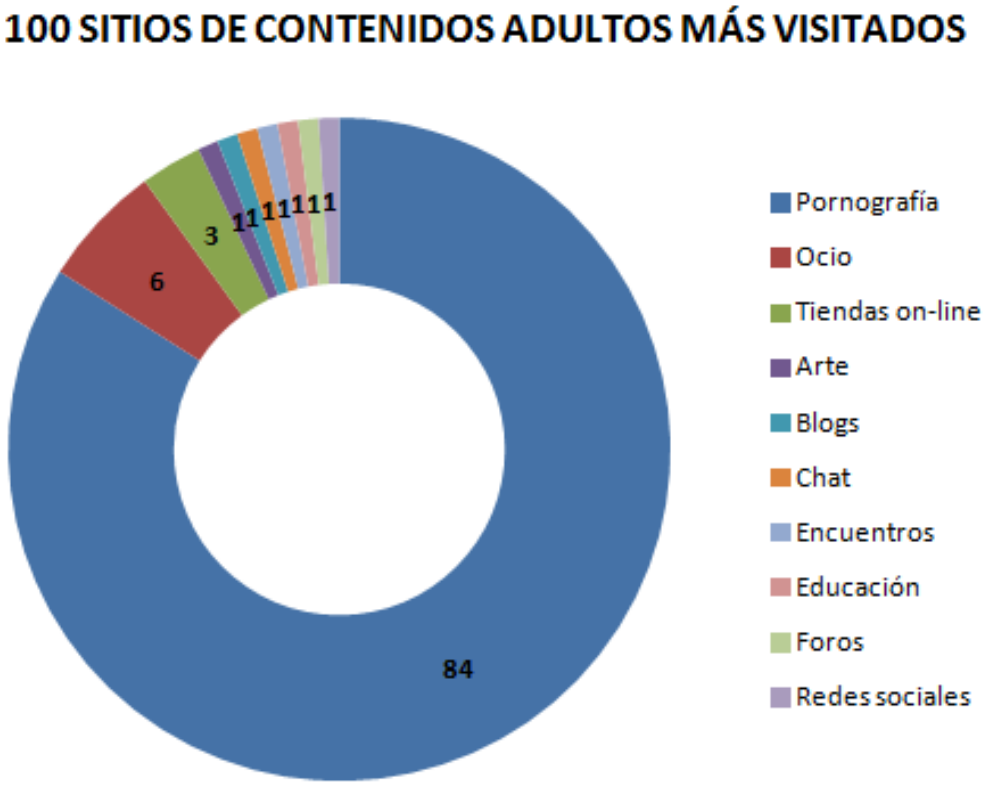


Ilustración 28. Análisis de los 100 sitios de contenido adulto más visitados

Como era de esperar, la pornografía es uno de los contenidos que más se visitan ya que hay grandes intereses económicos al ser uno de los negocios que más dinero mueve.

En esta lista, no se ha detectado ningún falso positivo y se puede ver como el ocio o las compras on-line siguen siendo unas de las temáticas más visitadas.

Alexa ofrece una lista con un total de 500 sitios dentro de esta categoría. Para analizar si los resultados siguen siendo igual de buenos, en cuanto al número de falsos positivos

y/o negativos, vamos a analizar a continuación los sitios situados en la parte más baja de su ranking, esto es, los situados entre las posiciones 400 y 500.

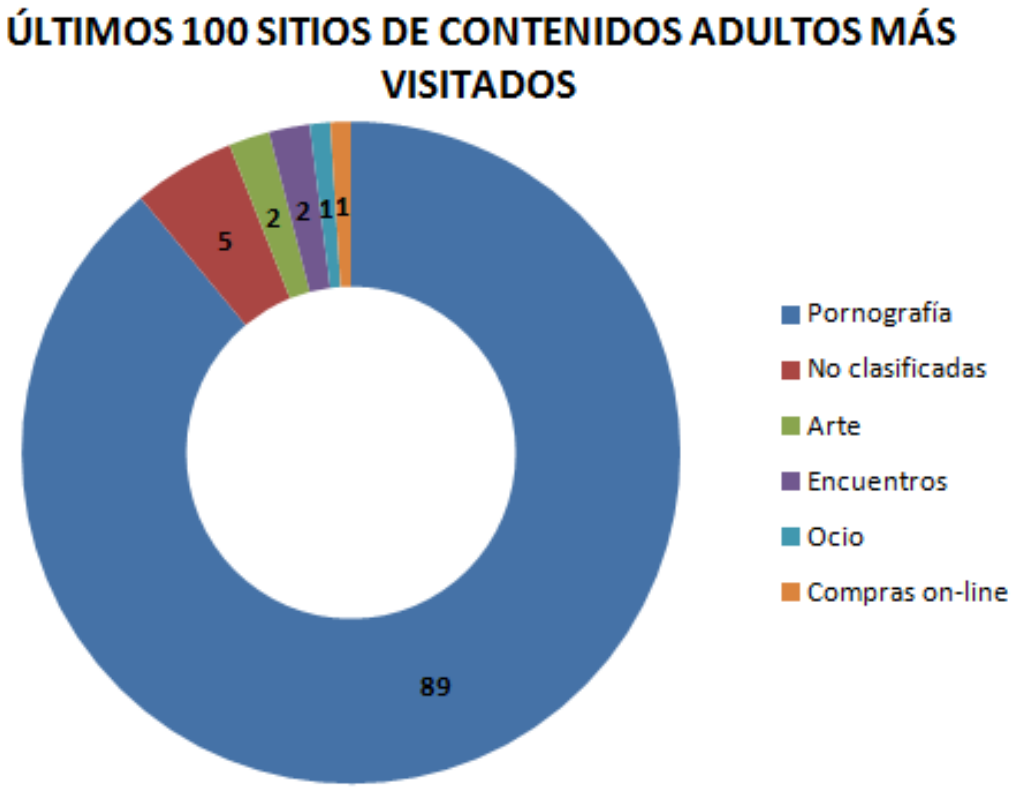


Ilustración 29. Análisis de los sitios de contenido adulto más visitados situados entre las posiciones 400 y 500

En este caso la efectividad del filtro ha bajado notablemente ya que de éstos últimos 100 sitios, un total de 5 no han devuelto ningún contenido y que al revisarlas deberían haber sido categorizadas como pornografía.

Con respecto al resto de clasificaciones, no se detectaron falsos positivos y volvemos a ver quela pornografía es el contenido más visitado.

2.3.4.3. Top 100 en España

Alexa también presenta estadísticas de navegación según los países. Así analizaremos el contenido de los 100 sitios más visitados en España.

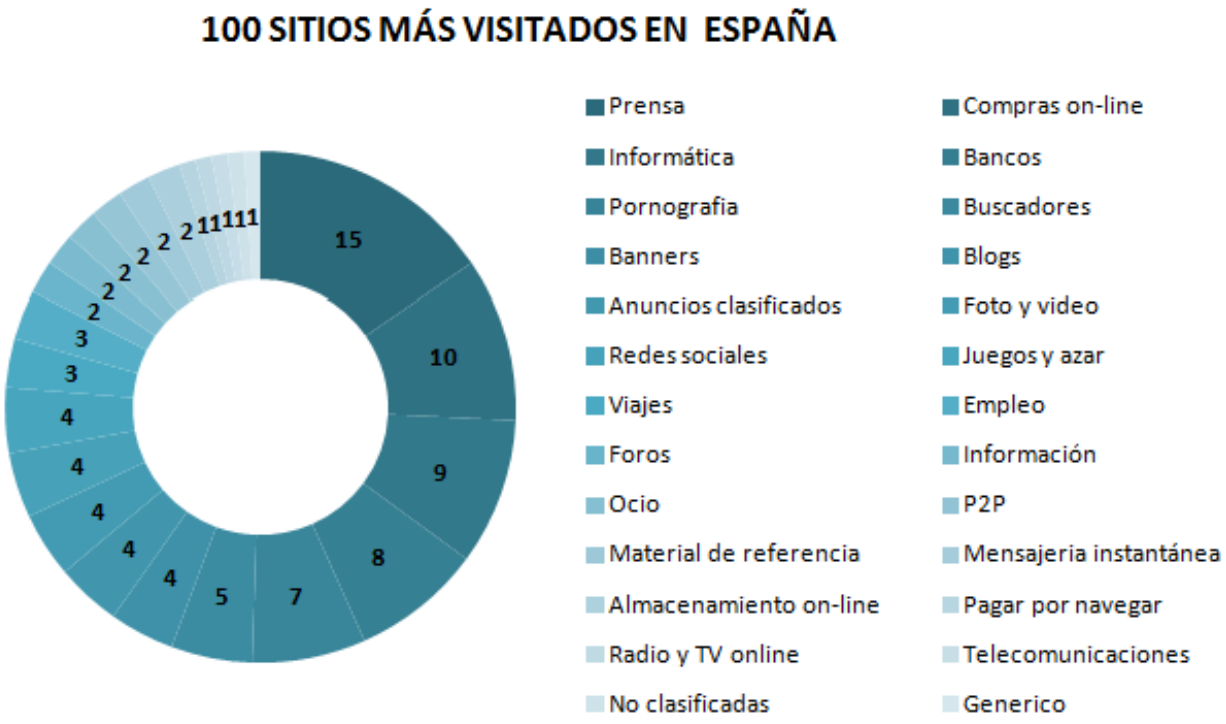


Ilustración 30. Análisis de los 100 sitios más visitados en España

En España, vemos que aunque la pornografía se encuentra entre los contenidos más visitados en mayor medida los españoles están más interesados por leer la prensa, realizar compras on-line, acceder a sus bancos o la informática.

Aunque los buscadores también son otros de los sitios más visitados, sorprendentemente las redes sociales aparecen tímidamente. Señalar también que en España vemos cómo *juegos y azar* (por ejemplo, *Loteriasyapuestas.es*) aparece como nueva categoría ya que esta temática está ampliamente extendida en la población.

De nuevo, no se ha detectado ningún falso positivo pero si dos falsos negativos o URLs no clasificadas (*Outbrain.comuser y Googleadservices.com*).

CAPÍTULO 4: Análisis funcional

1. Introducción

En este capítulo veremos un ejemplo de aplicación de un sistema de filtrado de contenidos con funcionalidades similares a las vistas en un posible entorno real.

Comenzaré diseñando el sistema y haré una estimación tanto de los tiempos de implantación del proyecto como de sus costes asociados.

2. Hábitos del uso de Internet

Para dar un mayor énfasis a la necesidad de las soluciones de filtrado de contenidos y, por consiguiente, a la razón de este proyecto veremos a continuación un gráfico en el que muestro de forma estadística el número de usuarios de Internet durante el año 2014 generado a partir de los datos presentados por la ITU.

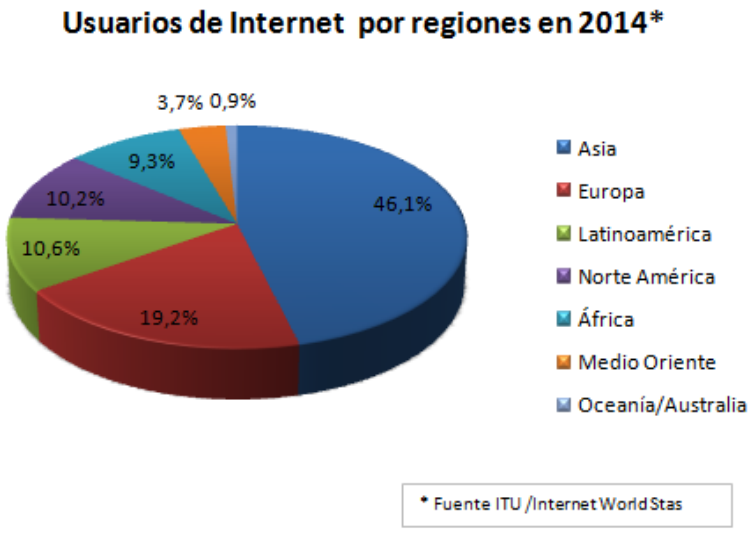


Ilustración 31. Usuarios de Internet por regiones en 2014

Según la ITU, a finales de 2014, el número de usuarios de Internet en todo el mundo habría alcanzado casi los 3.000 millones. Lo que correspondería a una penetración de usuarios de Internet del 40% a nivel mundial, situándose el 78% en los países desarrollados y el 3% en los países en vías de desarrollo.

A partir de los datos proporcionados por la ITU se ha generado el siguiente mapa mundial con la penetración de Internet según regiones:

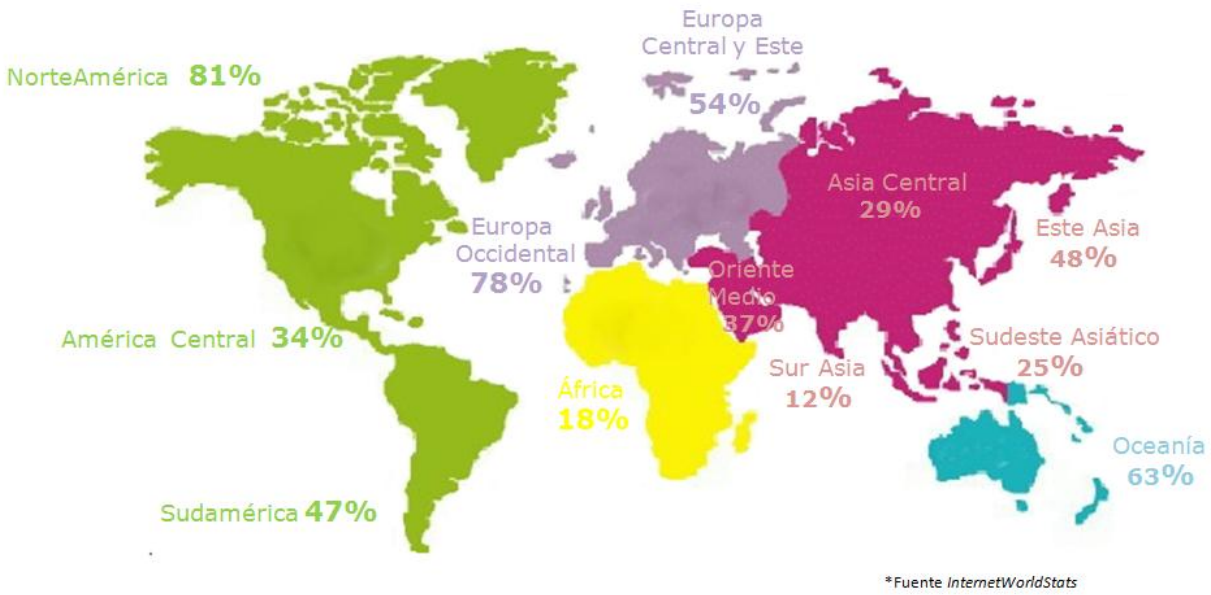


Ilustración 32. Penetración de Internet en enero de 2014

En África, vemos que casi el 20% de la población estaba en línea a finales de 2014, cuando en el año 2010 apenas era del 10%.

En las Américas, se estimaba que casi dos de cada tres personas usarían Internet a finales de 2014, lo que representaba la segunda mayor tasa de penetración después de la de Europa.

En Europa, la penetración de Internet alcanzaría el 75% (es decir, tres de cada cuatro personas) a finales de 2014 y sería la más alta a nivel mundial.

Un tercio de la población de Asia y el Pacífico estaría conectada a finales de 2014 y cerca del 45% de los usuarios de Internet totales procederían de esta región.

Este uso apremiante de Internet ha dispuesto que en la actualidad exista una sensibilidad general sobre el control de contenidos ilegales en Internet. Numerosos países

tienen implementadas según la legislación vigente soluciones de filtrado de contenidos basadas en:

- Proporcionar un filtrado legal para todos los usuarios a partir de listas proporcionadas por los mismos gobiernos.
- Ofrecer un servicio de filtrado a ciertos colectivos como menores y/o adolescentes.

En países de Asia, Oriente Medio y África existen este filtrado es ejercido directamente por el gobierno en condiciones muy restrictivas y que están fuera del propósito de este proyecto.

3. Problemática

Centrándonos en España, en el año 2014 la penetración de Internet estaba en torno al 77%. Se había reducido notablemente la penetración de la telefonía fija, cuyo descenso durante los últimos cinco años fue comprensible debido al gran uso de la telefonía móvil.

Pero, dependiendo del segmento de análisis, tanto la distribución del tráfico como los dispositivos de conexión utilizados son muy diferentes.

Así en un entorno empresarial, el tráfico predominante sería HTTP y HTTPS. Semejante a un entorno móvil en el que la mayoría de las conexiones estarían basadas en tráfico HTTP y HTTPS. Mientras que en un entorno residencial, el tráfico P2P podría encontrarse a la par que el HTTP y HTTPS.

Para el caso de uso elegido para este proyecto, el entorno elegido será un segmento empresarial con las siguientes características:

- Número de usuarios: 6000

- Ancho de banda: 400 Mbps
- Tipo de dispositivos: ordenadores con sistema operativo Windows 7, Windows Vista y Windows 8.
- Instalación en alta disponibilidad.
- Integración con directorios externos (LDAP, Directorio Activo de Microsoft).
- Mantenimiento de 24 meses de los elementos hardware y software.
- Solución de filtrado de contenidos para tráfico HTTP con:
    - Bloqueo de contenidos según categorías
    - Bloqueo de páginas web no categorizadas si muestran contenidos no adecuados.

4. Diseño de la solución

4.1. Topología de la solución

El modo de despliegue de la solución de filtrado de contenidos web será proxy.

4.2. Identificación y autenticación de usuarios

La función principal del sistema de filtrado es permitir o denegar el acceso a los contenidos web solicitados. Para ello necesita analizar la página web y en base al resultado de la detección y otra serie parámetros como, por ejemplo, qué usuario está solicitando dicha página, permitir o denegar el acceso.

Para determinar acciones en base al usuario que solicita un contenido, éste tiene que ser identificado mediante algún parámetro como su dirección IP, nombre de usuario o MSISDN dándose así a conocer en el sistema.

Pero desde el punto de vista de seguridad, se podrá completar la identificación con la autenticación o verificación de la identidad del usuario. La autenticación podrá ser externa o interna dependiendo de si es necesario acceder o no a otra plataforma para verificar la identidad del usuario. Si la autenticación es externa, la comunicación entre los sistemas deberá estar securizada para evitar cualquier tipo de suplantación de identidad.

Aunque existen diferentes métodos de autenticación según los parámetros que se utilicen en la verificación, para nuestro sistema sólo tendrá sentido una autenticación basada en un dato conocido como un password o credencial. El proceso general de autenticación constará de los siguientes pasos:

1. Para el acceso a Internet, el usuario necesita tener validado el acceso.

2. El sistema solicita al usuario que se autentique.

3. El usuario proporciona las credenciales necesarias para su identificación y autenticación.

4. El sistema valida, según sus reglas, si las credenciales aportadas son suficientes para permitir el acceso al usuario o no.

De acuerdo a la descripción del entorno en el que se quiere desplegar el sistema de análisis de contenidos web, la solución propuesta deberá integrarse con una base externa a través de consultas LDAP o *Lightweight Directory Access Protocol.* La información obtenida de dicha base de datos será el usuario y, opcionalmente, el grupo o grupos de usuarios a los que pertenece.

Mediante autenticación básica, el navegador siempre solicitará al usuario que se autentique. Esa información la recibirá el proxy a través de un campo en la cabecera HTTP y ese usuario quedará autenticado si el proxy es capar de obtener el mismo usuario del servidor externo con el que se comunica vía LDAP.

```
CONNECT www.google.es:443 HTTP/1.0
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko
Host: www.google.es
Content-Length: 0
DNT: 1
Proxy-Connection: Keep-Alive
Pragma: no-cache
Proxy-Authorization: Basic dXNlckBBVVRIMjoxMjM0NTY=

..,.........U.....q~e..-.u.......q.K~.....". 6.p.O'8....r[..D......$...<......4.
(.'.............,.+.$.#.
...=.<.5./.j.@.8.2.
.........N........
www.google.es..........
..................
.........................................4wJO.;.4I....O..:...yM....A.tp.....*.b#.L...%
\.....L....(..Q......y|.f.c>.......9v...d.!.....d0N....pZ...*../.M.:~.|9...`{}
3l.............(.........U...v2..?..L7E.
$=.y..B...=..a$
```

Ilustración 33. Petición HTTP con autenticación básica

Desde el punto de vista de seguridad, la autenticación básica no es robusta ya que se envían las credenciales codificadas en base64 (*Proxy-Authorization: Basic dXNlckBBVVRIMjoxMjM0NTY=*). Este tipo de codificación es muy sencilla de revertir a través de las numerosas webs posibilitan su decodificación instantánea.

Por tanto, sería recomendable utilizar otro tipo de autenticación que utilizase protocolos de comunicación más seguros como NTLM, *NT Lan Manager*, o Kerberos. En estos, la información de autenticación va cifrada mediante un token en la cabecera HTTP y dependiendo del tipo de autenticación, se podrá obtener el usuario al momento, NTLM, o tras solicitar una clave privada, Kerberos.

4.3. Definición hardware y software

Al solicitar alta disponibilidad deberemos utilizar como mínimo dos servidores para alojar la solución de filtrado. En principio, el esquema de funcionamiento será en modo activo-pasivo.

En el apartado 2 del capítulo anterior, vimos que los requisitos mínimos necesarios para poder instalar la solución del filtrado eran:

- 1 interfaz de red
- 4GB RAM
- 20GB HDD

Por lo tanto, los servidores deberán cumplir estas mínimas especificaciones.

Vimos también que el rendimiento del proxy utilizado en la maqueta era inferior a 500 Mbps, pero la limitación no provenía de la solución de filtrado sino del entorno virtual. Asique podríamos suponer que el mismo software instalado en servidores físicos mejoraría el rendimiento de la maqueta. Por tanto, cada uno de los servidores soportaría el *troughput* máximo de 400 Mbps solicitado en este caso de uso.

4.4. Redirección del tráfico de usuario

Como ya hemos visto anteriormente, todos los usuarios de la empresa deberán configurar sus navegadores para enviar el tráfico a nuestro proxy.

Para realizar esta configuración de forma transparente a los usuarios, se utilizará un fichero denominado *proxy pac*. Así gracias a la configuración de este fichero se gestionará la alta disponibilidad de los servidores, atacando siempre a uno de ellos hasta que deje de responder para entonces atacar al otro.

A continuación veremos un ejemplo sencillo de un proxy pac:

```
function FindProxyForURL(url, host){

 var proxy_principal = "<IP_PROXY1>:<port>";   //Proxy Primario
 var proxy_secundario = "<IP_PROXY2>:<port>";  // Proxy Secundario
 var proxy_no = "DIRECT";                      //No se usa el proxy

 var myIP = myIPAddress();               //IP local
 var theHost = host.toLowerCase();       //Host solicitado
 var theHostIP = dnsResolve( theHost );  //IP del host solicitado
```

```
  // Local host y loopback nunca usarán el proxy
  if (( "localhost" == theHost ) ||( shExpMatch(theHost, "localhost.*" )) ||( "127.0.0.1"
== theHost )) {
          return proxy_no;
  }

  //Excepción por URL
  if (shExpMatch(url, "http://www.mycompanywebsite.com*")){
          return proxy_no;
  }

  //Excepción por dirección IP
  if (isInNet (myIP, <Ip>, <máscara>)) {
          return proxy_no;
}

  // DEFAULT: Uso del proxy en un escenario de failover
  return "proxy_principal; proxy_secundario";

}
```

Ilustración 34. Ejemplo de fichero proxy pac

5. Diagrama de Gant

El éxito de cualquier proyecto depende en gran medida del grado de entendimiento de la problemática del cliente, de que se presente una solución adecuada a sus condiciones particulares y del cumplimiento de los compromisos asumidos.

O lo que es lo mismo, para lograr la satisfacción en el desarrollo de un proyecto deben estar en el equilibro las tres dimensiones siguientes: el tiempo, el presupuesto (los recursos comprometidos) y el alcance (las especificaciones del producto o servicio a obtener).

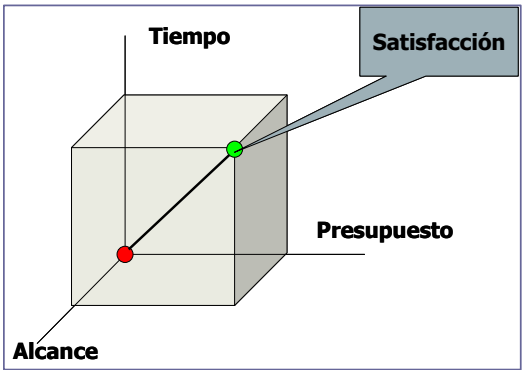


Ilustración 35. Dimensiones de un proyecto

Por ello, la metodología genérica de gestión de proyectos atiende a cinco grandes grupos de procesos definidos como sigue:

- **Proceso de inicio:** comprometer desde el inicio del proyecto y generar todo lo necesario para la consecución del éxito del mismo.
- **Proceso de planificación**: desarrollar y mantener un plan de trabajo viable para conseguir los objetivos del proyecto.
- **Proceso de ejecución:** coordinar al personal y a los recursos para seguir el plan desarrollado.
- **Proceso de control:** asegurar que los objetivos del proyecto sean alcanzados por medio de la monitorización y medición del progreso, así como la implementación de medidas correctivas cuando éstas sean necesarias.
- **Procesos de cierre:** formalizar la aceptación del proyecto, llevándolo a buen término.

Cada uno de los procesos anteriores estará formalizado por una serie de etapas detalladas en el siguiente diagrama de Gant:

	Task Name	Duration	Start	Finish	Predecessors
1	Análisis y diseño inicial de la solución propuesta	3 days	Tue 25/08/15	Thu 27/08/15	
2	Aprobación de los procedimientos	1 day	Tue 01/09/15	Tue 01/09/15	1
3	PO recibido	1 day	Wed 02/09/15	Wed 02/09/15	2
4	Cierre del diseño final de la arquitectura	1 day	Thu 03/09/15	Thu 03/09/15	3
5	**Hardware**	**8,75 days**	**Fri 04/09/15**	**Wed 16/09/15**	**4**
6	HW recibido por el suministrador del servicio	1 hr	Fri 04/09/15	Fri 04/09/15	
7	Instalación y configuración del SO	1 day	Fri 04/09/15	Mon 07/09/15	6
8	Configuración del direccionamiento de red	4 hrs	Mon 07/09/15	Mon 07/09/15	7
9	Instalación de la solución de filtrado	3 days	Mon 07/09/15	Thu 10/09/15	8
10	Envío del HW (2uds) al cliente	1 hr	Thu 10/09/15	Thu 10/09/15	9
11	Recepción del HW por ell cliente	2 days	Thu 10/09/15	Mon 14/09/15	10
12	Instalación del HW (enracado y cableado)	1 day	Mon 14/09/15	Tue 15/09/15	11
13	Configuración de la red del cliente	1 day	Tue 15/09/15	Wed 16/09/15	12
14	**Producción (proxy #1)**	**3,25 days**	**Wed 16/09/15**	**Mon 21/09/15**	**5**
15	Configuración de la solución de filtrado	2 hrs	Wed 16/09/15	Wed 16/09/15	5
16	Verificación de conectividad con el proxy #2	2 hrs	Thu 17/09/15	Thu 17/09/15	15
17	Desvío de tráfico de pruebas	3 hrs	Thu 17/09/15	Thu 17/09/15	16
18	Pruebas de aceptación	1 day	Thu 17/09/15	Fri 18/09/15	17
19	Validación del servicio en proxy #1	1 day	Fri 18/09/15	Mon 21/09/15	18
20	Desvío de tráfico real	3 hrs	Mon 21/09/15	Mon 21/09/15	19
21	**Producción (proxy #2)**	**3,38 days**	**Tue 22/09/15**	**Fri 25/09/15**	**14**
22	Configuración de la solución de filtrado	2 hrs	Tue 22/09/15	Tue 22/09/15	14
23	Verificación de conectividad con el proxy #1	2 hrs	Tue 22/09/15	Tue 22/09/15	22
24	Desvío de tráfico de pruebas	3 hrs	Tue 22/09/15	Tue 22/09/15	23
25	Pruebas de aceptación	9 hrs	Tue 22/09/15	Wed 23/09/15	24
26	Validación del servicio en proxy #2	1 day	Thu 24/09/15	Thu 24/09/15	25
27	Desvío de tráfico real	3 hrs	Fri 25/09/15	Fri 25/09/15	26
28	Formación	3 days	Fri 25/09/15	Wed 30/09/15	14;21
29	Baby sitting	5 days	Fri 25/09/15	Fri 02/10/15	14;21

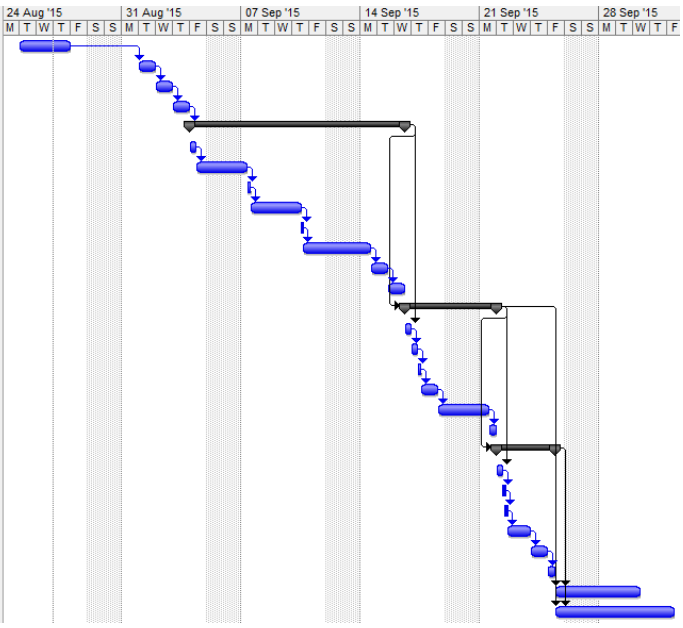


Ilustración 36. Diagrama de Gant

De acuerdo a la planificación anterior, en poco más de un mes se podría llevar a cabo la implantación de una solución de filtrado de contenidos bajo los requisitos definidos en el apartado 3 de este capítulo.

A continuación describiremos las fases que componen el plan de trabajo diseñado:

1. Análisis y diseño inicial de la solución propuesta

Durante el proceso de inicio, la figura del jefe de proyecto es fundamental ya que se encargará de analizar las necesidades del cliente y de realizar la estimación del esfuerzo necesario para llevar a cabo el proyecto determinando la solución que más se ajuste a sus requisitos y para ellos definirá una estrategia.

Posteriormente planificará el proyecto en todos sus aspectos, identificando las actividades a realizar, los recursos necesarios, los plazos y los costes de toda la ejecución.

La solución técnica será presentada al cliente mediante una reunión presencial.

2. Aprobación de los procedimientos

Tras acordar con el cliente lo detalles de la solución que se va a implementar, el jefe de proyecto generará la documentación pertinente donde se recojan los detalles de la misma.

Dicha documentación deberá ser aprobada por el cliente.

También se le facilitará al cliente, como parte de la documentación anterior, una oferta que recoja todos los costes derivados de la acometida del proyecto. Para abaratar costes, será también el jefe de proyecto quién generará la oferta y discutirá con el cliente los detalles de la misma.

3. PO recibido

Para poder comenzar con la ejecución del proyecto, es necesario recibir por parte del cliente el PO o *Purchase Order* asociado a la oferta anteriormente facilitada y aceptada por el cliente.

4. Cierre del diseño final de la arquitectura

El cliente facilitará los detalles necesarios de su arquitectura de red así como todos los datos del direccionamiento de red que se precisen para poder desplegar los proxys de la solución de filtrado (direcciones IP, máscaras de red, gateway…).

Este será el momento en que el jefe de proyecto proponga, en su caso, modificaciones a los límites u objetivos básicos del proyecto si concurriesen circunstancias que así lo aconsejasen.

Seguidamente, se alcanzaría la fase de ejecución que estaría siempre supervisada por el jefe de proyecto.

5. Hardware

Dentro de esta fase se englobarán todas las tareas necesarias para preparar el hardware previo a su instalación en el CPD o *Centro de Procesamiento de Datos* del cliente.

Una vez recibidos los servidores por el proveedor de la solución de filtrado de contenidos, es tarea del técnico de sistemas la instalación del sistema operativo y la configuración del direccionamiento IP así como de todos los datos necesarios para que se pueda alojar e instalar la solución.

El ingeniero de despliegue se encargará posteriormente de la instalación del software de la solución de filtrado de contenidos.

Tras finalizar estas tareas, ambos servidores serán enviados al cliente que se encargará de alojarlos en su CPD, disponiéndolos en los racks y cableándolos. Quedarían listos para ejecutar las verificaciones de red y validación de la solución.

6. Producción (proxy #1)

Esta fase comprendería todos los trabajos necesarios para que el primer servidor proxy quedase desplegado en producción como, por ejemplo, la integración de la autenticación de los usuarios o la configuración de las categorías permitidas y bloqueadas. Estos trabajos serán realizados por el Ingeniero de Despliegue.

Una vez configurada la solución de filtrado, se realizarán las primeras pruebas de validación de la integración en red del servidor así como de la funcionalidad de la solución.

Para poder realizar el plan de pruebas, se seguirá en detalle la documentación generada durante la entrega de los procedimientos en la fase 2. Entre otras, este plan de pruebas recogerá evaluaciones de la verificación de la correcta integración de la plataforma con el directorio activo o la validación de la autenticación de los usuarios. Por supuesto, es el momento de verificar la funcionalidad de la solución de filtrado.

Tras finalizar el plan de pruebas, se le entregará al cliente un informe con los resultados obtenidos para la aceptación y aprobación provisional de la solución desplegada en este primer servidor.

7. Producción (proxy #2)

Esta fase sería idéntica a la anterior realizada sobre el segundo servidor. Pero como diferencia, dentro del plan de pruebas habría que incluir la evaluación del correcto funcionamiento en alta disponibilidad del sistema.

Una vez acabadas estas pruebas, la puesta en producción de la solución de filtrado estaría finalizada.

8. Formación

Una vez que la solución de filtrado se encontrase totalmente desplegada en producción, se procederá a realizar la formación al personal que la empresa haya determinado como administradores de la plataforma.

El objetivo de la formación será que tras la implementación del servicio, el cliente no se encuentre con un escenario tecnológicamente desconocido. El formador se encargará de instruir al equipo de trabajo del cliente en la gestión y el *troubleshooting* de la nueva plataforma.

Se ha decidido realizar en este punto la formación, para familiarizar al personal encargado de gestionar el servicio de filtrado de contenidos con la propia plataforma en producción. Durante la formación se irán resolviendo las primeras dudas o incidencias relacionadas con la configuración del servicio de filtrado directamente en producción.

9. Baby sitting

Una vez que la plataforma proxy se encuentra en producción, comenzará el período de *baby sitting* o tiempo de observación de la solución que resulta esencial durante la etapa posterior a la implementación y la puesta en marcha del servicio de filtrado.

Tras finalizar éste, se le solicitaría al cliente la aprobación final de la solución de filtrado.

Comenzarían a contar el tiempo de servicio contratado, 24 meses.

6. Simulación de costes

En este apartado veremos el coste de la implantación de laa solución de filtrado de contenidos descrita a lo largo de este capítulo detallando el gasto de personal y de los elementos software y hardware que se generarían durante su desarrollo y tiempo de servicio.

6.1. Coste de personal

La jornada laboral que se ha tenido en cuenta en el diagrama de Gant del apartado anterior está compuesta de 8 horas.

Así, a partir de la planificación diseñada previamente, el coste de personal según el cómputo de horas dedicado al proyecto sería:

Perfil	Coste hora	Núm. horas	Importe
Jefe de Proyecto	100 €	128	12.800 €
Técnico de Sistemas	50 €	10	500 €
Ingeniero de Despliegue	100 €	25	2.500 €
Formador	50 €	24	1.200 €
		TOTAL	**17.000 €**

Tabla 5: Coste de personal

6.2. Costes de viajes y dietas

Para definir este coste se ha tenido en cuenta que el jefe de proyecto debería mantener al menos una reunión presencial con el cliente durante la fase de inicio. El resto de costes de viaje los incurrirían tanto el Ingeniero de Despliegue durante la fase de implantación como el formador durante la fase de formación.

Uds	Descripción	Precio unitario	Importe
3	Vuelos	200 €	600 €
8	Hotel	100 €	800 €
11	Dieta	80 €	880 €
		TOTAL	**2.280 €**

Tabla 6: Coste de viajes

6.3. Coste de software y licencias

El coste del servicio de filtrado se realizará en base al número de usuarios contratados.

Uds	Producto	Importe
6000	Filtrado de contenidos Web	54.000 €
	TOTAL	**54.000 €**

Tabla 7: Coste software

6.4. Coste de hardware

No se ha determinado ningún servidor concreto y únicamente se ha definido las características principales en cuanto a interfaces, disco duro y RAM de acuerdo a la maqueta sobre la que se realizaron las pruebas de latencia y rendimiento.

Uds	Producto	Importe
2	Servidor: 1U 1 Procesador Intel Xeon 4 GB RAM 160 GB SATA Tarjeta de Red Intel Dual-Port Gigabit	5.000 €
	TOTAL	**10.000 €**

Tabla 8: Coste hardware

6.5. Coste del soporte y mantenimiento de la solución

Para este coste se ha tenido en cuenta la duración del servicio definida por el cliente: 24 meses.

Años	Descripción	Importe
2	Soporte 5x8	30.000 €
2	Actualizaciones	40.000 €
	TOTAL	**70.000 €**

Tabla 9: Coste de soporte y mantenimiento

6.6. Coste total del proyecto

A continuación se detalla el coste total del proyecto teniendo en cuenta todos los costes anteriores.

Descripción	Costes totales
Personal	17.000 €
Viajes	2.280 €
Software	54.000 €
Hardware	10.000 €
Soporte y mantenimiento	70.000 €
TOTAL	**153.280 €**

Tabla 10: Coste total

A este coste habría que añadir un coste del 21% en concepto de impuesto sobre el valor añadido o IVA.

Descripción	Costes total
Total sin IVA	153.280 €
IVA (21%)	32.139 €
TOTAL	**185.469 €**

Tabla 11: Coste total con IVA incluido

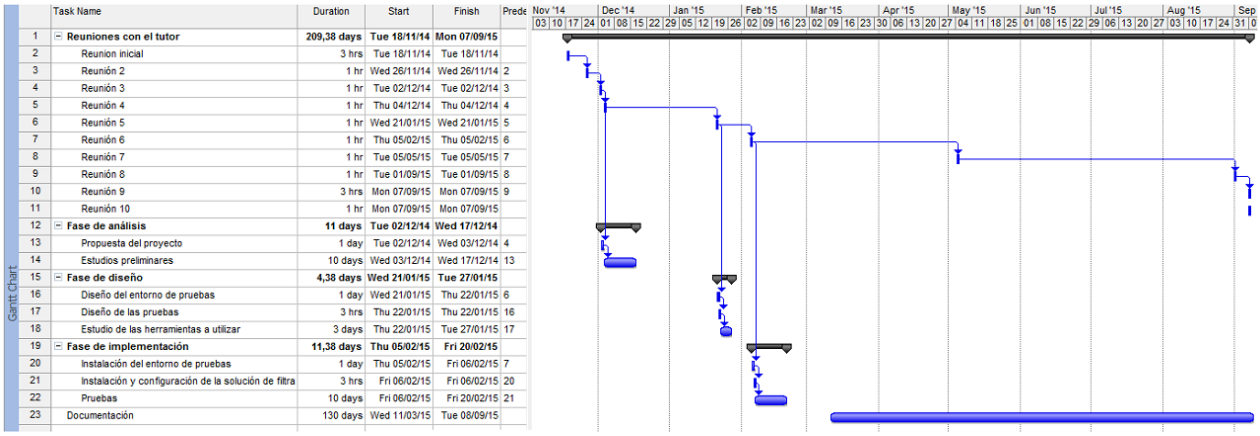
CAPÍTULO 5: Presupuesto del PFC

1. Introducción

En este capítulo veremos los tiempos empleados en cada una de las actividades que han conllevado este Proyecto Fin de Carrera y se realizará un estudio de los costes asociados al mismo.

2. Diagrama de Gantt

A continuación se presenta en el siguiente diagrama de Gant la planificación de las tareas y tiempos estimados que se siguieron a lo largo de este Proyecto Fin de Carrera.



	Task Name	Duration	Start	Finish	Prede
1	**Reuniones con el tutor**	**209,38 days**	**Tue 18/11/14**	**Mon 07/09/15**	
2	Reunion inicial	3 hrs	Tue 18/11/14	Tue 18/11/14	
3	Reunión 2	1 hr	Wed 26/11/14	Wed 26/11/14	2
4	Reunión 3	1 hr	Tue 02/12/14	Tue 02/12/14	3
5	Reunión 4	1 hr	Thu 04/12/14	Thu 04/12/14	4
6	Reunión 5	1 hr	Wed 21/01/15	Wed 21/01/15	5
7	Reunión 6	1 hr	Thu 05/02/15	Thu 05/02/15	6
8	Reunión 7	1 hr	Tue 05/05/15	Tue 05/05/15	7
9	Reunión 8	1 hr	Tue 01/09/15	Tue 01/09/15	8
10	Reunión 9	3 hrs	Mon 07/09/15	Mon 07/09/15	9
11	Reunión 10	1 hr	Mon 07/09/15	Mon 07/09/15	
12	**Fase de análisis**	**11 days**	**Tue 02/12/14**	**Wed 17/12/14**	
13	Propuesta del proyecto	1 day	Tue 02/12/14	Wed 03/12/14	4
14	Estudios preliminares	10 days	Wed 03/12/14	Wed 17/12/14	13
15	**Fase de diseño**	**4,38 days**	**Wed 21/01/15**	**Tue 27/01/15**	
16	Diseño del entorno de pruebas	1 day	Wed 21/01/15	Thu 22/01/15	6
17	Diseño de las pruebas	3 hrs	Thu 22/01/15	Thu 22/01/15	16
18	Estudio de las herramientas a utilizar	3 days	Thu 22/01/15	Tue 27/01/15	17
19	**Fase de implementación**	**11,38 days**	**Thu 05/02/15**	**Fri 20/02/15**	
20	Instalación del entorno de pruebas	1 day	Thu 05/02/15	Fri 06/02/15	7
21	Instalación y configuración de la solución de filtra	3 hrs	Fri 06/02/15	Fri 06/02/15	20
22	Pruebas	10 days	Fri 06/02/15	Fri 20/02/15	21
23	Documentación	130 days	Wed 11/03/15	Tue 08/09/15	

Ilustración 37. Diagrama de Gant del PFC

Tras las reuniones con el tutor para determinar el temario y delimitar el contenido del PFC, se realizó una notable labor de documentación para poder conocer tanto los orígenes como las transformaciones que el análisis de contenido ha sufrido desde sus inicios hasta la actualidad. Así como las diferentes tecnologías de filtrado.

Otro punto destacable en el que se dedicó parte del mayor esfuerzo fue durante el despliegue del entorno de pruebas y realización de las pruebas cuyos resultados vistos anteriormente.

3. Presupuesto

En este caso, el presupuesto que se presenta en este capítulo engloba el coste estimado en personal, elementos de hardware y software empleado durante la realización de este Proyecto Fin de Carrera.

3.1. Coste de personal

Para el cómputo de horas, tendremos en cuenta que la jornada laboral en este caso se vió reducida a 5 horas (media estimada). Así, a partir de la planificación presentada en el anterior diagrama de Gant, el coste de personal necesario para la realización de este PFC sería:

Perfil	Coste hora	Núm.horas	Importe
Analista	50 €	11	550 €
Diseñador	50 €	5	250 €
Técnico	45 €	12	540 €
Responsable de documentación	20 €	130	2.600 €
		TOTAL	**3.940 €**

Tabla 12: Coste de personal PFC

3.2. Coste de software y licencias

En este apartado se incluyen las licencias de las diferentes aplicaciones y herramientas utilizadas durante la realización del PFC.

Descripción	Coste imputable
Microsoft Office Word 2007	60 €
Microsoft Office Power Point 2007	
Microsof Office Excel 2007	
Microsoft Office Project 2007	
Solución filtrado de contenidos	80 €
VMWare EXSi4	250 €
Apache Software	0 €
Notepad+++	0 €
TOTAL	**390 €**

Tabla 13: Coste software y licencias PFC

3.3. Coste de hardware

En este caso tendremos en cuenta los equipos empleados durante el despliegue de la maqueta necesaria para la evaluación del sistema y la redacción de la memoria.

Producto	Precio	Período de Depreciación	Tiempo de uso (meses)	Coste Imputable
Portátil Lenovo Thinkpad R400	800 €	60	5,23	70 €
Ratón y teclado Logitech	80 €	60	5,23	70 €
TFT DELL 1708FPt	50 €	60	5,23	70 €
Impresora HP Deskjet F2180	49 €	60	1	13 €
			TOTAL	**222 €**

Tabla 14: Coste hardware PFC

3.4. Coste total del proyecto

A continuación se detalla el coste total del proyecto teniendo en cuenta todos los costes anteriores.

Descripción	Costes totales
Personal	3.940 €
Software	390 €
Hardware	222 €
TOTAL	**4.552 €**

Tabla 15: Coste total PFC

Añadiendo al coste anterior el 21% en concepto de impuesto sobre el valor añadido o IVA, obtendríamos:

Descripción	Costes total
Total sin IVA	4.552 €
IVA (21%)	956 €
TOTAL	**5.508 €**

Tabla 16: Coste total PFC con IVA incluido

CAPÍTULO 6: Conclusiones y futuros trabajos

Tras el estudio realizado en este proyecto de un sistema de análisis de contenidos web, expondré las conclusiones que obtenidas y las futuras líneas de trabajo.

1. Conclusiones

El filtrado de contenidos es una de las piedras angulares de la seguridad en Internet. Pero una solución como la analizada en este proyecto, no es una solución sencilla de mantener.

En primer lugar, como hemos visto el sistema de filtrado de contenidos está basado en listas de categorías y técnicas de aproximación paramétrica. Por tanto, su efectividad vendrá dada por lo bien categorizadas que estén las páginas web y lo bien entrenados que estén los modelos que definan el sistema de decisión.

Tímidamente vimos al utilizar la información provista por Alexa sobre sitios más visitados, que la solución de filtrado mostraba ciertos dominios que no estaban clasificados, denominados como "falsos negativos". En este punto, una relación bidireccional entre el proveedor de la solución de filtrado y cliente se haría necesaria para que cuando un usuario detectase algún dominio o URL no clasificado, inmediatamente lo pusiese en conocimiento del proveedor y así éste puede realizar su clasificación y actualización de sus bases de datos.

No obstante, el problema más importante de este tipo de sistemas de análisis de contenidos son los "falsos positivos" causados por errores en la clasificación de los datos que afectan a los resultados de la detección. Pero para solucionar estos errores de clasificación de este tipo, también será necesario implementar algún tipo de herramienta o mecanismo que los corrija porque este tipo de errores podrían conllevar demandas legales de los proveedores de los sitios que están siendo improcedentemente bloqueados. Tras las

pruebas realizadas con los dominios proporcionados por Alexa, la solución de filtrado utilizada denotaba una alta efectividad en la clasificación de los dominios.

Referente a las listas de URLs englobadas por categorías, este tipo de sistemas exigen algún tipo actualización periódica tanto de los nuevos sitios que vayan apareciendo como de las contribuciones realizadas por los usuarios.

Por otro lado, aunque apenas es apreciable desde el punto de vista de la experiencia de usuario, los sistema de filtrado ralentizan la comunicación entre el usuario y el servidor web al introducir ciertas tareas de cómputo operacional.

El rendimiento obtenido de la solución utilizando una máquina virtual no era muy bueno ya que se desaprovecha más del 50% del volumen de trabajo que, en principio, era capaz de gestionar el sistema. Esta limitación venía dada por el entorno virtual utilizado asique el uso de servidores dedicados siempre ofrecerá mejores resultados.

El sistema descrito no es tangible. Hasta que un usuario no es bloqueado, no es consciente que entre él e Internet hay 'algo'. Pero si no hay bloqueo, no hay forma de confirmar que la solución está funcionando. Así una forma sencilla de mostrar que el análisis de contenidos está actuando, sería la implementación de algún tipo de herramienta que logase las acciones que está llevando a cabo para componer algún tipo de informe de uso que los usuarios o administradores pudiesen consultar.

Aunque no se ha analizado en este proyecto, ya que no era base de estudio, la herramienta de gestión del sistema de filtrado deberá estar diseñada de forma intuitiva y sencilla.

2. Futuras líneas de investigación

Otro aspecto a considerar en el proceso de filtrado, sería la posibilidad de inclusión de listas blancas y/o negras. Estas listas contendrían URLs individuales a las cuales se

permitiría o bloquearía, respectivamente, la navegación. Con ello se podría ofrecer la oportunidad a los usuarios finales o administradores de los sistemas de validar o invalidar en base a ciertos criterios, los contenidos existentes independientemente de la valoración del sistema de filtrado. Pero es muy importante conocer el criterio utilizado para generar esta definición de estas listas para no anular el sentido de la solución de filtrado.

Muchos de los contenidos de página web de Internet están basados en imágenes, en este caso, la solución de filtrado que hemos visto resultaría ineficaz. Por tanto, otra de las posibles líneas de investigación que se podrían seguir sería la implementación del análisis de imágenes. Aunque actualmente existen algunas soluciones como la que proporciona el gigante Google llamada *Safe Search*, pero es una medida restringida ya que en este caso sólo se pueden filtrar las imágenes devueltas por una búsqueda de Google.

La solución vista en este proyecto, está orientada a mitigar usos imprudentes o inadecuados de navegación web. Sin embargo, y como vimos en los capítulos iniciales, el uso de Internet no se reduce únicamente a consultas de páginas web. Por ello, analizando desde el punto de vista de seguridad nuestro sistema de filtrado, otra serie de protecciones que podrían incluirse serían:

- Antivirus para proteger a los usuarios finales de los virus, troyanos, gusanos, botnets…que actualmente inundan la red.

- Análisis de protocolos para inspeccionar en detalle el contenido de los ficheros que se intercambian los usuarios, por ejemplo, en las redes P2P.

- Análisis de correos electrónicos ofreciendo protección frente a spam, virus, phishing o incluso la reputación de las empresas o proveedores de servicios de Internet impidiendo que sus direcciones IP entrasen en listas negras debido a un uso inadecuado de sus usuarios.

GLOSARIO DE TÉRMINOS

AFCC	American Federal Communication Commission
ARPA	Advanced Research Projects Agency
ARPANet	Advanced Research Projects Agency Network
CPD	Centro de Procesamiento de Datos
DNS	Domain Name System
DPI	Deep Packet Inspection
FOSI	Family Online Safety Institute
GW	Gateway
HLD	High Level Design
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ICRA	Internet Content Rating Association
IMP	Interface Message Processor
IP	Internet Protocol
IRC	Internet Relay Chat
ISO	International Organization for Standarization
ITU	International Telecommunication Union
IVA	Impuesto sobre el Valor Añadido
LDAP	Lightweight Directory Access Protocol
NAPT	Network Address and Port Translation
NCP	Network Control Protocol
NPL	National Physical Laboratory
NT	New Technology
NTLM	NT Lan Manager
OVF	Open Virtual Machine Format
PC	Personal Computer
PFC	Proyecto Fin de Carrera
PO	Purchase Order
PPS	Peticiones Por Segundo
P2P	Peer-to-peer
RSAC	Recreational Software Advisory Council
RTC	Red Telefónica Conmutada
SOAP	Simple Object Access Protocol
TCP	Transfer Control Protocol

TLD Top Level Domain

URL Uniform Resource Locator

WWW World Wide Web

Categoría Tipo de contenido que puede tener una página web o cualquier información susceptible de ser categorizada.

Clasificación Determinación de si el contenido de un texto es positivo en una categoría determinada.

Conjunto de entrenamiento Conjunto de ficheros de entrada que se utiliza para el entrenamiento. Se necesitan dos conjuntos, uno de ficheros que son positivos y otro que son negativos para el tipo de diccionario y la categoría que se quiere entrenar. Cuanto mejores son los conjuntos de entrenamiento, mejor podrá salir el diccionario. También es necesario que los conjuntos sean suficientemente grandes para que el entrenamiento sea capaz de extraer la información necesaria de ellos y generar un buen diccionario. Por esta razón, en ocasiones es conveniente replicar un número determinado de veces un conjunto de entrenamiento (utilizar los ficheros varias veces), aunque al repetirse los ficheros, dependiendo de la correcta clasificación de los mismos y el tipo de diccionario que se está entrenando podría falsear los resultados y obtener un diccionario de peor calidad. Normalmente suelen utilizarse conjuntos de entrenamiento como ficheros agrupados para realizar los entrenamientos.

Detección Proceso de la aplicación por el cual se determina si un texto (contenido a analizar) da positivo o no en una determinada categoría. Básicamente consiste en la suma de los pesos de un diccionario según las palabras que contenga el texto a analizar (esta suma se realiza una sola vez por contenido, es decir, que si una determinada palabra aparece varias veces en el texto, sólo se sumará el peso una vez), y si el resultado supera el umbral se considera que el texto da positivo en dicha categoría.

Diccionario Herramienta fundamental para la detección de categorías por contenido. Es un fichero que contiene un conjunto de palabras asociadas a un peso. Es el resultado de los entrenamientos.

Eficacia Porcentaje de aciertos de un diccionario sobre un conjunto que debe dar positivo. A mayor eficacia, mejor es el diccionario.

Entrenamiento Proceso por el cual, dados dos conjuntos de entrenamiento, uno positivo y otro negativo, se genera un diccionario.

Filtro web Aplicación que, en base a ciertas herramientas, determina si una página web puede ser visualizada o no según la configuración establecida.

Palabra Unidad mínima sin contenido en una comunicación.

Peso Valor numérico normalmente asociado a las palabras del diccionario, creado durante el entrenamiento y utilizado en la detección, que indica la ponderación que tendrá esa palabra en referencia a la categoría analizada.

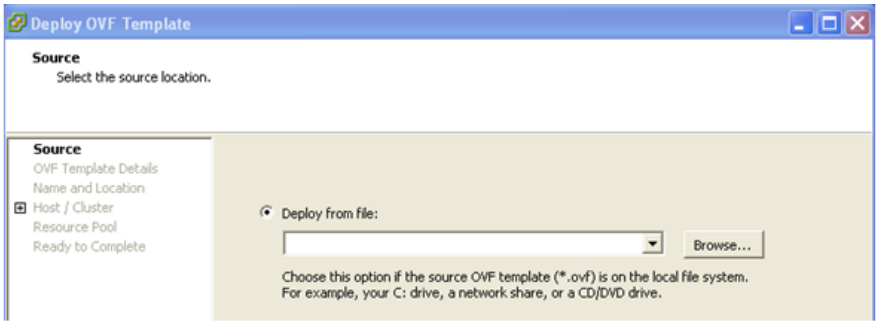
Umbral Valor límite sobre el cual se considera que una detección es positiva y negativa si es inferior al mismo. Por coherencia, se suele utilizar el mismo tanto en el entrenamiento como en la detección.

ANEXO A: Instalación y configuración del entorno virtual VMWare ESXi 4

Instalación

Conectar el cliente VmWare vSphere a nuestro entorno ESXi:

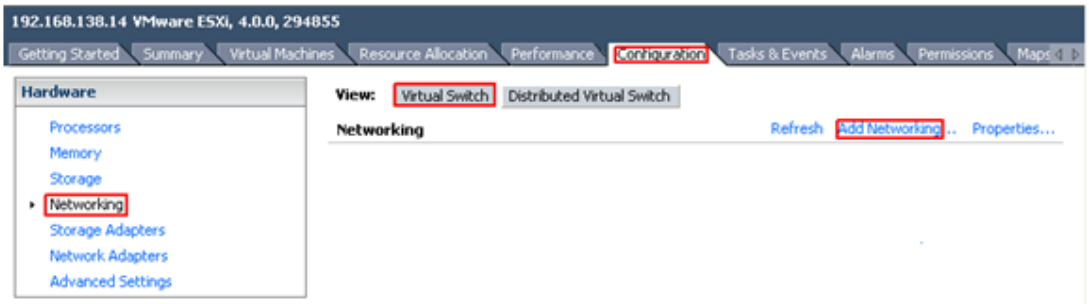
- Seleccionar *File* -> *Deploy OVF Template*. Añadir el fichero .ovf con la imagen del appliance virtual (http://exinda.force.com/virtual)



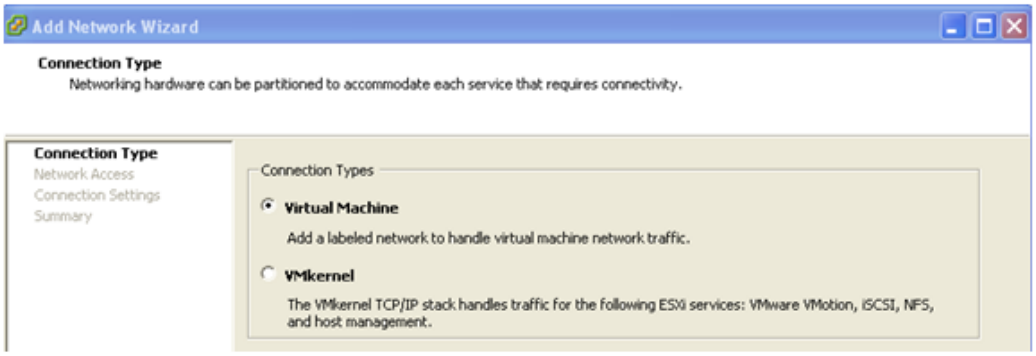
- Completar los pasos de la instalación:

Configuración

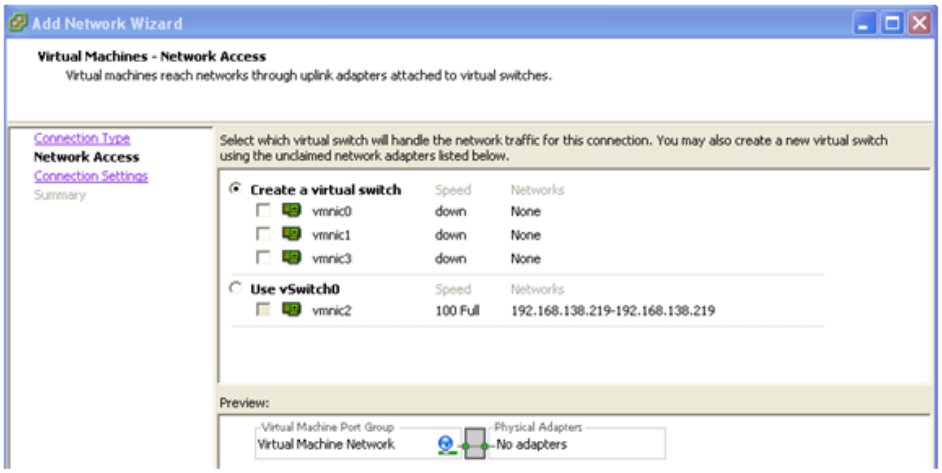
1. Desde vSphere, seleccionar el host desde el panel de inventario y hacer click en la pestaña *Configuration* > *Networking.* Seleccionar *Virtual Switch*:



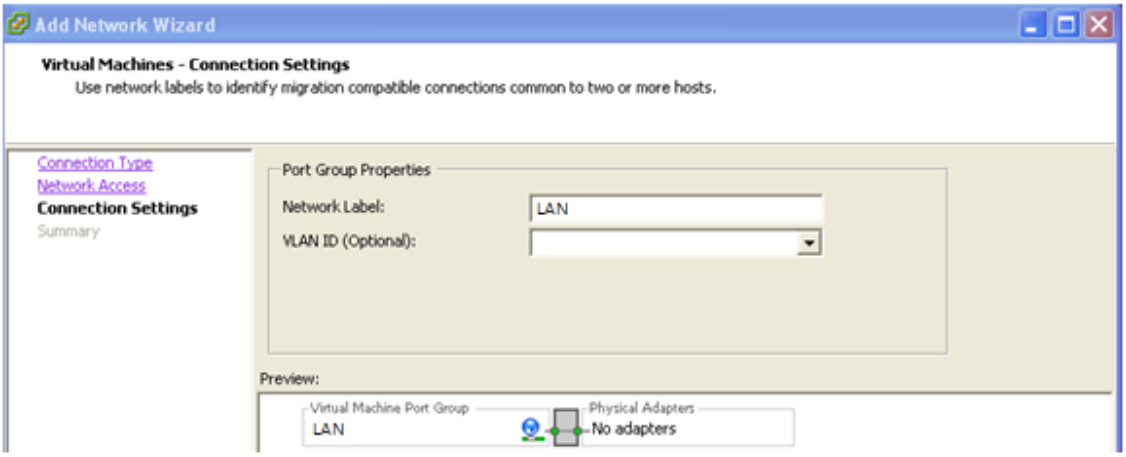
2. Hacer click en *Add Networking*. Seleccionar la opción Virtual Machine y hacer click en Next.



3. Seleccionar *Create a virtual switch* y hacer click en *Next*.

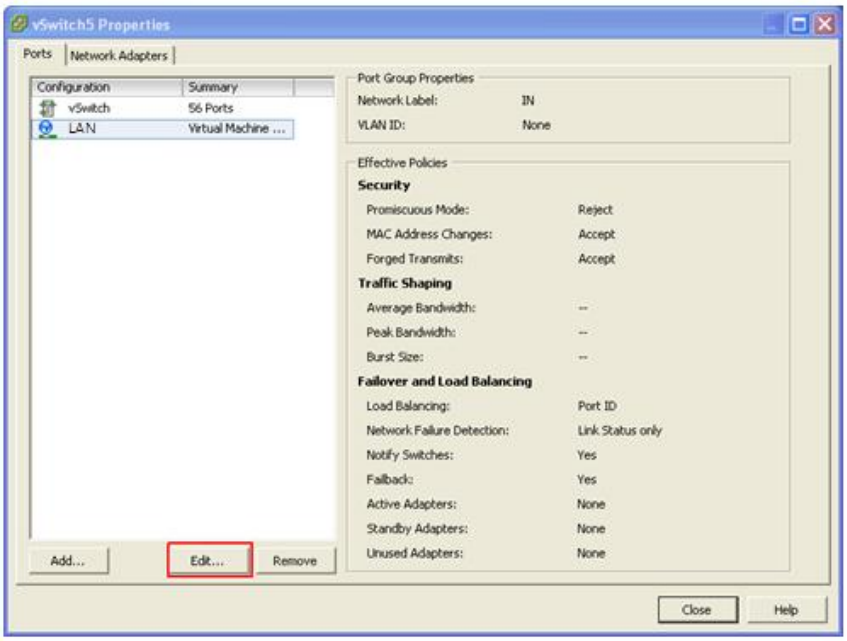


4. Dentro de *Port Group Properties* introducir un nombre de red (LAN). Hacer click en *Next*. Y después en *Finish*.

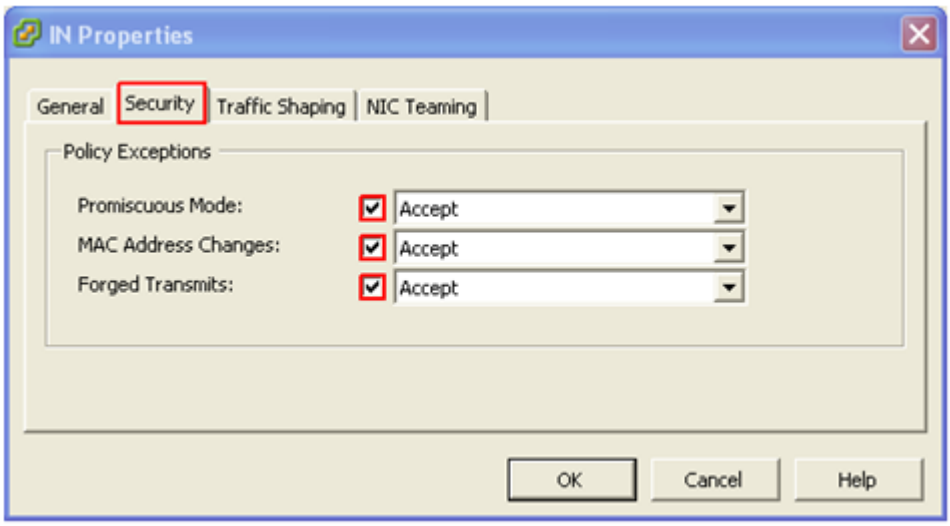


5. Seleccionar el switch virtual creado y hacer click en *Properties*.

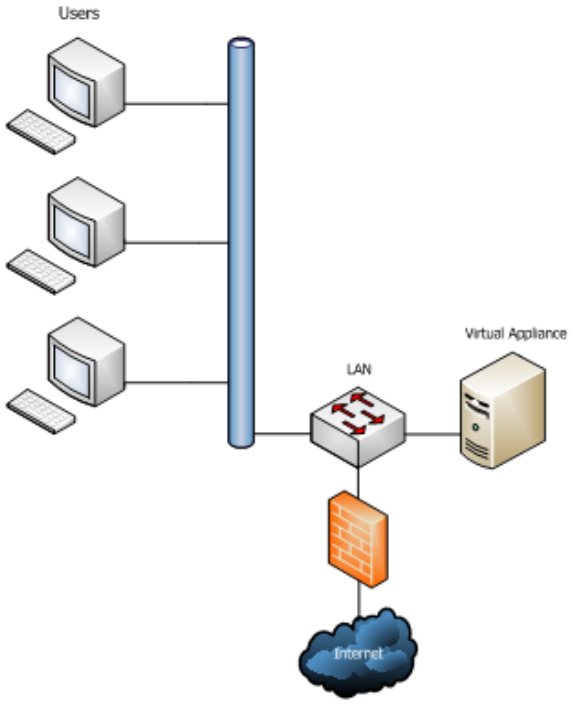
6. Hacer click en *Edit* y editar su configuración por defecto.



7. En la pestaña Security, habilitar Policy Exceptions y seleccionar para cada una de las opciones que aparecen. Hacer click en OK.



Una vez que tenemos el switch virtual creado y configurado, hay que configurar las interfaces del appliance virtual para desplegarlo en modo proxy.



8. Seleccione la VM desde el panel *Inventory*. Hacer click en *Edit*.

9. El appliance virtual presenta cuatro adaptadores de red:

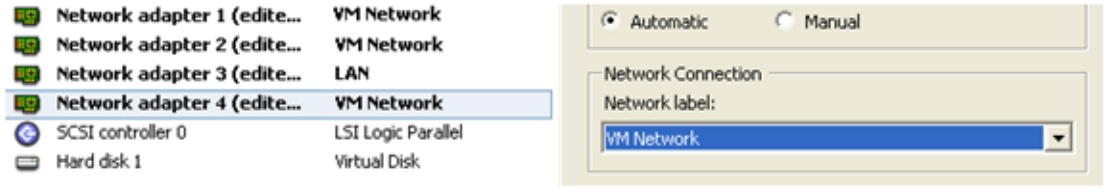
   Network adapter 1 is eth0

   Network adapter 2 is eth1

   Network adapter 3 is eth2

   Network adapter 4 is eth3

   Para desplegar de acuerdo a una topología proxy, los adaptadores de red se han de configurar de la siguiente manera:



   El tráfico ha de llegar al switch virtual LAN.

10. Por último, encender la máquina virtual.

ANEXO B: Código script generador de peticiones

```
#!/bin/bash

# $1 fichero con la lista de URLs

while read URL
do
    # Lanzamos las peticiones a través del proxy
    curl -x <IP_PROXY>:<PUERTO_PROXY> -m 5 $URL &

    # Lanzamos las peticiones directamente a Internet
    #   curl -m 5 $URL &

        #Esperamos hasta que el curl acabe
        echo -ne "$j $i  \r"
        conn=$(ps -ef |grep curl |grep -v 'grep' |wc -l)
        #Definimos el número de peticiones #100
        while [ $conn -gt 100 ]
        do
            sleep 0.1
            conn=$(ps -ef |grep curl |grep -v 'grep' |wc -l)
        done

done < $1
```

ANEXO C: Listados de sitios de Alexa

TOP 100 MUNDIAL		TOP 100 ESPAÑA	
SITIO	CATEGORIZACIÓN	SITIO	CATEGORIZACIÓN
Google.com	Search engines	Google.es	Search engines
Facebook.com	Social networks	Google.com	Search engines
Youtube.com	Photo and video	Facebook.com	Social networks
Baidu.com	Search engines	Youtube.com	Photo and video
Yahoo.com	Search engines	Amazon.es	Shopping
Amazon.com	Shopping	Twitter.com	Social networks
Wikipedia.org	Reference material	Live.com	Computing
Qq.com	Portals	Yahoo.com	Search engines
Twitter.com	Social networks	Wikipedia.org	Reference material
Google.co.in	Search engines	Marca.com	Press
Taobao.com	Leisure	Aliexpress.com	Shopping
Live.com	Web mail	Elpais.com	Press
Sina.com.cn	Portals	Elmundo.es	Press
Yahoo.co.jp	Search engines	Milanuncios.com	Classified ads
Linkedin.com	Employment	T.co	Computing
Weibo.com	Social networks	Linkedin.com	Employment
Ebay.com	Shopping	Wordpress.com	Blogs
Google.co.jp	Search engines	Instagram.com	Social networks
Yandex.ru	Search engines	Ebay.es	Shopping
Vk.com	Social networks	As.com	Press
Hao123.com	Search engines	Msn.com	Shopping
Blogspot.com	Blogs	Lacaixa.es	Financial institutions
T.co	Computing	Paypal.com	Financial institutions
Bing.com	Search engines	Booking.com	Travel
Google.de	Search engines	Xvideos.com	Pornography
Instagram.com	Social networks	Segundamano.es	Classified ads
Aliexpress.com	Shopping	Abc.es	Press
Msn.com	Shopping, Portals	Googleadservices.com	NOT CATEGORIZED
Amazon.co.jp	Shopping	Pinterest.com	Social networks
Google.co.uk	Search engines	Outbrain.comuser	NOT CATEGORIZED
Reddit.com	Social networks	Idealista.com	Classified ads
Ask.com	Search engines	Onclickads.net	Banners
Google.com.br	Search engines	Elconfidencial.com	Press
Pinterest.com	Social networks	Amazon.com	Shopping
Onclickads.net	Banners	Eltiempo.es	Info
Google.fr	Search engines	Pubted.com	Banners
Microsoft.com	Computing	Forocoches.com	Forum
Wordpress.com	Blogs	Tripadvisor.es	Travel

Tmall.com	Shopping	Wordreference.com	Reference material
Paypal.com	Economy	Microsoft.com	Computing
Mail.ru	Portals	Tumblr.com	Blogs
Sohu.com	Portals	Infojobs.net	Employment
Tumblr.com	Blogs	Bbva.es	Financial institutions
Imgur.com	Photo and video	Gruposantander.es	Financial institutions
Google.ru	Search engines	Apple.com	Computing
Xvideos.com	Pornography	Ingdirect.es	Financial institutions
Imdb.com	Leisure	Elcorteingles.es	Shopping
Apple.com	Computing	Blogger.com	Blogs
Fc2.com	Computing	Pornhub.com	Pornography
Google.it	Search engines	Bing.com	Search engines
Google.es	Search engines	Movistar.es	Telecommunications
Googleadservices.com	NOT CATEGORIZED	Lavanguardia.com	Press
Netflix.com	Photo and video	20minutos.es	Press
Amazon.de	Shopping	Mundodeportivo.com	Press
360.cn	Computing	Bancsabadell.com	Financial institutions
Stackoverflow.com	Computing	Ask.com	Search engines
Tianya.cn	Forum	Fotocasa.es	Classified ads
Craigslist.org	Classified ads	Sport.es	Press
Alibaba.com	Shopping	Bet365.es	Gambling
Ok.ru	Social networks	Minube.com	Travel
Google.com.mx	Search engines	Eleconomista.es	Press
Google.ca	Search engines	Rtve.es	Press
Pornhub.com	Pornography	Telecinco.es	Radio and tv online
Google.com.hk	Search engines	Xhamster.com	Pornography
Diply.com	Leisure	Mejortorrent.com	P2P servers
Naver.com	Search engines	coches.net	Shopping
Amazon.co.uk	Shopping	Pccomponentes.com	Computing
Gmw.cn	Press	Stackoverflow.com	Computing
Rakuten.co.jp	Shopping	Elitetorrent.net	P2P servers
Xhamster.com	Pornography	Bankia.es	Financial institutions
Go.com	Search engines	Expansion.com	Press
Blogger.com	Blogs	Dropbox.com	Online storage
Kat.cr	P2P servers	Loteriasyapuestas.es	Gambling
Adcash.com	Banners	Bongacams.com	Pornography
Outbrain.com	Generic	Imdb.com	Leisure
Cnn.com	Press	Aemet.es	Info
Soso.com	Search engines	Filmaffinity.com	Leisure
Nicovideo.jp	Photo and video	Blogspot.com	Blogs
Google.com.tr	Search engines	Flickr.com	Photo and video
Amazon.in	Shopping	Tubecup.com	Pornography

Flipkart.com	Shopping	Coches.net	Shopping
Xinhuanet.com	Press	Hootsuite.com	Computing
Cntv.cn	Press	Ikea.com	Shopping
Google.co.id	Search engines	Bancosantander.es	Financial institutions
Booking.com	Travel	Mediamarkt.es	Shopping
People.com.cn	Press	Adcash.com	Banners
Bbc.co.uk	Press	Directrev.com	Generic
Github.com	Computing	Whatsapp.com	Instant messaging
Pixnet.net	Portals	Adf.ly	Pay per surf
Googleusercontent.com	Search engines	Libertaddigital.com	Press
Google.com.au	Search engines	Pasion.com	Pornography
Google.co.kr	Search engines	Pordede.com	Photo and video
Dropbox.com	Online storage	Imgur.com	Photo and video
Google.pl	Search engines	Popads.net	Banners
Ebay.de	Shopping	Redtube.com	Pornography
Popads.net	Banners	Lavozdegalicia.es	Press
Dailymotion.com	Photo and video	Taringa.net	Forum
Espn.go.com	Press	Feedly.com	Computing
Livedoor.jp	Portals	Softonic.com	Computing
Ebay.co.uk	Shopping	Bet365.com	Gambling

TOP 100 CONT.ADULTO		ÚLTIMOS 100 CONT.ADULTO	
SITIO	CATEGORIZACIÓN	SITIO	CATEGORIZACIÓN
Xnxx.com	Pornography	Bbs.mediumpimpin.com	Pornography
Youporn.com	Pornography	Annas-angels.co.uk	Pornography
Livejasmin.com	Pornography	Tgp.89.com	Pornography
G.e-hentai.org	Pornography	Cuckoldvideoclips.com	Pornography
Adultfriendfinder.com	Pornography	Sinfulcelebs.freesexycomics.com	Pornography
Flirt4free.com	Pornography	Britishsexcontacts.com	Pornography
Cam4.com	Pornography	Gayasianams.com	Pornography
Nudevista.com	Pornography	Nextdoormale.com	Pornography
Xcams.com	Pornography	Sandyssuperstars.com	Pornography
Fetlife.com	Pornography	Targetescorts.com	Pornography
Adam4adam.com	Pornography	Brutalasia.com	Pornography
Freeones.com	Pornography	Spankstories.com	Pornography
Literotica.com	Pornography	Sexpo.com.au	Pornography
Ebaumsworld.com	Leisure	Lukeisback.com	Pornography
Playboy.com	Pornography	Cartoonvalley.com	Pornography
Manhunt.net	Pornography	Swingersboard.com	Pornography
Planetsuzy.org	Pornography	Stripclublist.com	Pornography

Imlive.com	Pornography	Paramour.com.au	Pornography
Furaffinity.net	Art	Allasiandvd.com	Pornography
Newgrounds.com	Leisure	Babes6.com	Pornography
Digitalplayground.com	Pornography	Spankingblog.com	Pornography
Payserve.com	Pornography	Angel-elite-escort.com	Pornography
Asexstories.com	Pornography	Mistressalexia.com	Pornography
Clips4sale.com	Pornography	Femalecompanions.com	Pornography
Mrskin.com	Pornography	Lovehentaimanga.com	Pornography
Nhentai.net	Pornography	Ultimate3dporn.com	Pornography
Fakku.net	Pornography	Babylongirls.co.uk	Pornography
Oglaf.com	Pornography	Tpe.com/~altarboy/	Pornography
Shooshtime.com	Pornography	Legaction.com	Pornography
Asstr.org	Pornography	Sextoys.co.uk	Pornography
Streamate.com	Pornography	Kamasutra.com	Pornography
Fling.com	Pornography	Malespank.net	Pornography
Aebn.net	Pornography	Escortmadeira.com	Pornography
Iafd.com	Pornography	Pissblog.com/blog/	Pornography
Aventertainments.com	Pornography	Taratainton.com	Pornography
Voyeurweb.com	Pornography	Dcup.com	Pornography
Thehun.net	Pornography	Bgeast.com	Pornography
Cams.com	Pornography	Just18.com	Pornography
Squirt.org	Pornography	Wasteland.com	Pornography
Hentai-foundry.com	Pornography	Pierresilber.com	Shopping
Celebritymoviearchive.com	Leisure	Femalecelebrities.com	Pornography
Luscious.net	Pornography	Thisis.delvecomic.com/NewWP/	NOT CATEGORIZED
Worldsex.com	Pornography	Earlmiller.com	Pornography
Nifty.org	Pornography	Pornhome.com	Pornography
Ftvgirls.com	Pornography	Blissbox.com	Pornography
Videosexarchive.com	Pornography	Publicflash.com	Pornography
Fabswingers.com	Pornography	Cuckold-chastity-belt-stories.com	Pornography
Girlfriendvideos.com	Pornography	Smutnetwork.com	Pornography
Scoreland.com	Pornography	Adultwork.co.uk	Pornography
Private.com	Pornography	Drunkcyclist.com	Pornography
Vintage-erotica-forum.com	Pornography	Swingvillage.com	Pornography
Adultdvdempire.com	Pornography	Trottla.net	NOT CATEGORIZED
Suicidegirls.com	Pornography	Mattsmodels.com	Pornography
Adameve.com	Pornography	Serious-coin.com	NOT CATEGORIZED
Somethingpositive.net	Leisure	Slavefarm.com	Pornography
Femjoy.com	Pornography	Dollmate.jp	NOT CATEGORIZED
Indiansexstories.net	Pornography	Jadedvideo.com	Pornography
Swinglifestyle.com	Pornography	Straightfellas.com	Pornography
Dudesnude.com	Pornography	The-femdom.com	Pornography

Videobox.com	Pornography	Adultgamereviews.com	Pornography
Recon.com	Pornography	Australian-babe.com	Pornography
Peachyforum.com	Pornography	Lovercash.com	Datings
Lovehoney.co.uk	Pornography	665leather.com	Pornography
Adultdvdtalk.com	Pornography	Thepinupfiles.com	Leisure
Tonybatman.com	Pornography	Saafe.info	Pornography
F-list.net	Chat	Justsayah.com	Art
Joerogan.net	Blogs	Dsdoll.us	NOT CATEGORIZED
Www74.virtuagirl.com	Pornography	Mymasturbation.com	Pornography
Debonairblog.com/blog/	Pornography	Real-femdom.com	Pornography
Wickedweasel.com	Shopping	Fisting.com	Pornography
1999.co.jp/eng/	Shopping	Nudecelebsmagazine.com	Pornography
Niteflirt.com	Pornography	Diaper-bois.com	Pornography
Ma3comic.com	Leisure	Hentai-top100.com	Pornography
Odloty.pl	Pornography	Cdgirls.com	Pornography
Mplstudios.com	Pornography	Shemalestrokers.com	Pornography
Sdc.com	Datings	Thedvdcompany.com	Pornography
Gamelink.com	Pornography	Blonde-escorts-uk.co.uk	Pornography
Hentairules.net	Pornography	Makepure.com	Pornography
Manjam.com	Social networks	Aphrodite-agency.com	Pornography
Hustler.com	Pornography	Dirty-david.com	Pornography
Jlist.com	Shopping	Webmasters.asiamoviepass.com	Pornography
Https://inkbunny.net/	Forum	Bigdoggie.net	Pornography
Onlydudes.com	Pornography	Loverspackage.com	Pornography
Teendreams.com	Pornography	Shadowlane.com	Pornography
Penthouse.com	Pornography	Acemassage.net	Pornography
Rentboy.com	Pornography	Cherrygirls.co.uk	Pornography
Killsometime.com	Photo and video	Daddyswap.com	Pornography
Fleshlight.com	Pornography	Pavelphoto.com	Pornography
Wtfpeople.com	Pornography	Classyangel.com	Pornography
Buttsmithy.com	Pornography	Darkwanderer.net	Pornography
Rabbitsreviews.com	Pornography	The-clitoris.com	Sexuality
Southern-charms.com	Pornography	Us.vclart.net/vcl/	Art
Goaskalice.columbia.edu	Education	Smittenkittenonline.com	Pornography
Gaydemon.com	Pornography	Lovedreamer.com	Pornography
Sucksex.com	Pornography	Wyldesites.com	Pornography
1by-day.com	Pornography	Fuckforforest.com	Pornography
Lushstories.com	Pornography	Tantrachair.com	Pornography
Xbiz.com	Pornography	Koalaswim.com	Pornography
Tlavideo.com	Pornography	Gayromeo.com	Datings
Mcstories.com	Pornography	Photoclubs.com	Pornography

BIBLIOGRAFÍA

Berelson, B. (1952). *Content Analysis in Comunication Research*, Free Press, Glencoe.

Krippendorff, K.(1990). *Metodología de análisis de contenido. Teoría y Práctica.* Piados Comunicación.

Bardin, L. (1996 2ª e) *Análisis de contenido.* Akal.

Pool, I de Sola (1959) . *Trend in Content Analysis*, Urbana. University of Illinois Press.

Stone, P.J. Dunphy D.C, Smith M.S. Ogilvie D. M. (Ed) (1966), *The general inquirier. A computer approach to content analysis in the behavioural sciences.* Cambridge. Mas MIT Press.

Ruiz Olabuénaga, J.I. (1996). Metodología de la investigación cualitativa. Deusto.

Tanenbaum, Andrew S., Redes de computadoras, Pearson Educación(2003)

www.vmware.com

www.itu.int

httpd.apache.org

www.alexa.com